DOCUMENT RESUME

ED 152 825 TM 007 002

AUTHOR	Bossone, Richard M., Ed.; Weiner, Max, Ed.
TITLE	Proceedings. The National Conference on Testing: Major Issues. November 17 & 18, 1977.
INSTITUTION	City Univ. of New York, N.Y. Center for Advanced Study in Education.
REPORT NO	CASE-01-78
PUB DATE	78
NOTE	178p.
AVAILABLE FROM	CUNY Research Foundation, Dr. Richard M. Bossone, Center for Advanced Study in Education, Graduate School of CUNY, 33 West 42nd Street, New York, New York 10036 ($5.00)
EDRS PRICE	MF-$0.83 HC-$10.03 Plus Postage.
DESCRIPTORS	Ability Grouping; Achievement Tests; College Admission; Conference Reports; Educational Accountability; Educational Legislation; *Educational Testing; Elementary Secondary Education; Intelligence Tests; Legal Problems; *Legal Responsibility; *Minority Groups; Nature Nurture Controversy; Parent Reaction; Performance Factors; Post Secondary Education; Quotas; Standardized Tests; *Teacher Associations; Teacher Responsibility; Test Bias; *Testing Problems; *Test Interpretation; Test Results; Test Validity; Unions
IDENTIFIERS	Minimal Competency Testing

ABSTRACT

Educational testing and its uses have come under increasing criticism as they are used more and more for decision making--to determine student tracking assignments, college admission, personnel selection, and licensing policies. The papers from this conference, which discussed the major issues in testing, include: (1) Testing: A Parent's Point of View; (2) Must Educators be Lawyers Too?; (3) Legal Implications of Intelligence and Competency Testing; (4) Alternatives to Quotas; (5) Critical Concerns in the Testing of Minorities: Time for initiatives; (6) Testing Minority Children: Why, How, and with What Effects?; (7) Why Standardized Tests Fail the Accountability Test; (8) Testing; Teacher Unions and Professional Responsibility; (9) Testing and Teacher Unions; (10) Testing Uses and Misuses in Education as Related to Minimal Competency Testing; (11) On the Uses and Misuses of Standardized Tests in Education; and (12) Use and Abuse of Tests in Education. The summarizing speaker suggested that tests are here to stay, and raised the following issues which have not yet been resolved: suitable additional methods of evaluation; governance of testing programs; innate group differences affecting test scores; determination of cutting scores; fair tests for all populations; and open admissions policies. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152825

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Richard M. Bossone
Max Weiner

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM007 002

Proceedings

The National Conference on Testing: Major Issues

November 17 & 18, 1977

Edited by
Richard M. Bossone
Max Weiner

Center for Advanced Study in Education
The Graduate School and University Center
of the City University of New York

CASE 01-78
1978

TABLE OF CONTENTS

Page

INTRODUCTION 1

OPENING ADDRESS

Testing: What in the World Are We Arguing About?
Ellis B. Page 3

PART I TESTING AND LEGAL IMPLICATIONS

Testing: A Parent's Point of View
Blanche Lewis 16

Must Educators Be Lawyers Too?
Anthony J. Polemeni . . 23

Legal Implications of Intelligence and Competency Testing
Merle Steven McClung . . 29

PART II TESTING OF MINORITIES

Alternatives to Quotas
Richard B. Darlington. . 45

Critical Concerns in the Testing of Minorities: Time for New Initiatives
Ronald J. Samuda 57

Testing Minority Children: Why, How, and with What Effects?
Sandra Scarr 71

PART III TESTING AND TEACHER UNIONS

Why Standardized Tests Fail the Accountability Test
Frances Quinto 103

Testing: Teacher Unions and Professional Responsibility
Irwin H. Polishook . . . 116

Testing and Teacher Unions
Robert L. Ebel 125

Page

PART IV TESTING USES AND MISUSES IN EDUCATION

Testing Uses and Misuses in Education as Related to Minimal Competency Testing
Thomas W. Mann 138

On the Uses and Misuses of Standardized Tests in Education
E. Belvin Williams . . . 147

Use and Abuse of Tests in Education
Elizabeth Hagen 156

A SUMMARY STATEMENT
Ellis B. Page 168

SELECTED BIBLIOGRAPHY 171

PROCEEDINGS OF THE NATIONAL CONFERENCE ON TESTING: MAJOR ISSUES

INTRODUCTION

On November 17 and 18, 1977, the Center for Advanced Study in Education (CASE) of the Graduate School and University Center of the City University of New York (CUNY) sponsored a "National Conference on Testing: Major Issues." The Conference was held at the CUNY Graduate School, 33 West 42nd Street in New York City and was attended by approximately 100 people from all over the United States. The Conference Chairman was Dr. Richard M. Bossone, Professor of English at the CUNY Graduate School.

The main purpose of this Conference was to highlight and clarify major issues in testing, a topic of pressing concern to the entire educational community. Thirteen speakers presented papers on such topics as testing and legal implications, testing of minorities, testing and teacher unions, and testing uses and misuses in education; in addition, a forum was held which allowed those in the audience to ask questions and share their thinking on the variety of complex problems that relate to testing.

Tests and their use have come under increasing criticism in the last few years. The more tests are used in decision making, the more they are criticized. Tests are increasingly relied upon by educators and employers to determine what track students will follow in schools, who will be admitted to specific colleges, who will be employed, promoted or fired, or who will be licensed in a given profession. It is not surprising, therefore, that instruments so crucial in determining who will succeed in our society have strong defenders and equally strong detractors.

The papers in this publication discuss the major issues centering around testing. Some defend tests strongly, others moderately, one not at all. But whatever the authors' views, the reader will gain a better and more sophisticated understanding of the issues as argued by nationally recognized experts.

OPENING ADDRESS

TESTING: WHAT IN THE WORLD ARE WE ARGUING ABOUT?

Ellis B. Page
Professor of Educational Psychology
University of Connecticut

Giving the opening address at a meeting such as this has drawbacks and advantages. As you might have noticed, testing is, to put it gently, controversial. In controversies, it is sometimes quite exciting to be one of the debaters, to take a position in strong opposition to another. As much as one might enjoy the dust and smoke of battle, the opening address forces one to hold off a bit, a little distance from the battlefield. And perhaps from this perspective -- temporarily less tactical and more strategic -- we can ask the question: What in the world are we arguing about?

First, we are arguing about the explicit general topics of the conference: legal implications, minorities, teacher unions, and a category of "uses and abuses." These may appear to be four separate questions, but they are only partially so. We will see if we cannot notice some factors that in fact bind them all together.

At another level, we are arguing about rules of evidence. At still a third level, we are arguing about a desired shape of human society. And fourth, we are arguing about man's model of himself. In short, when we debate the current issues of testing, we may appear to be talking technical matters, but we debate against a background of the deepest and most perennial questions that perplex and delight humankind. We are arguing about a great deal "in the world."

Should Testing Be Abolished?

An elementary charge very much in the news is that testing should be "abolished." Those who argue for the abolition of testing have commanded a certain amount of public attention, and their arguments have, for the popular imagination, the great virtue of simplicity. Consider this quote, from a long-time opponent, Banesh Hoffman (Houts, 1975, p. 33): "Multiple choice tests penalize the deep student, dampen creativity, foster intellectual dishonesty, and undermine the very foundations of education." And as most people know, a major teachers' union has called for a "moratorium" on aptitude, ability, and achievement tests. But for pure extremity, it is hard to beat this quote by Jerrold Zacharias: "I feel emotionally toward the testing industry as I would toward any other merchants of death. I feel that way because of what they do to the kids. I'm not saying they murder every child -- only 20% of them." (quoted in Rudman, 1977, p. 183).

In other words, testing should simply disappear. It is a beautiful reduction like that made by groups who would abolish all taxes. Now there is an idea we could all subscribe to -- except for that tiny portion of the tax pie which supports our own endeavors! Yet no one -- no responsbile person -- takes seriously the no-tax demand. Should we be more serious about the no-testing demand?

Clearly, there are enemies of testing, some of whom have been identified as friends: humanists, psychotherapists, school people, social innovators, and philosophical Marxists (Page, April, 1977). On the other hand, there are many friends of testing. These include, obviously, those who sell tests. There are also: psychometricians, who have studied measurement; administrators and planners, who want to know the effects of programs; behavioral scientists, who want to understand human behavior; teachers, who want to gather objective outside evidence about the abilities and progress of their students; admissions and employment officers, who want to cut through the "old-boy" system of elite schools and friendly recommendations; and students themselves, who seek comparable and reliable evidence of their ability to succeed in educational and occupational programs.

The friends of testing seem to outnumber the enemies many times over. And when Frank Stetz investigated attitudes of administrators and board members, those most directly charged with governing the schools, there was no consensus favorable to test abolition (as one might infer from some organizations). To the contrary, even on a question asking whether tests "discriminate against minorities," 43% denied it. And when asked whether it would be "wise to eliminate standardized testing," these decision makers were in agreement by 86%. For their part, they clearly suspected (around 70% of them) that "Teacher groups (e.g., NEA) who call for elimination of standardized testing are only trying to avoid being held accountable" (quoted in Rudman, 1977, p. 184).

In short, despite the continuous drum-beat in favor of eliminating testing, abolition is truly not a threat. It is, rather, an emotional plaint, like the dislike for taxes, or technology, or traffic. At times, all of us must feel like Luddites and want to destroy the machine which seems to be replacing us. Yet, with the exceptions of saints and mystics, we find our pre-scientific innocence, once lost, seems forever beyond our reach.

Is Testing Growing Still Faster?

Recently there has been a startling development in Connecticut. In November, 1977, the Connecticut Board of Education voted a much wider use of high-school equivalency testing than has ever been the case. In the past, to take the equivalency exams a youth had to be 18 years old and out of school. Then, as an "adult," he could try to demonstrate that he had reached an acceptable standard over the content of an ordinary high-school curriculum. In fact, he had to make a considerably better showing than many of his age-mates who graduated in the usual manner: He had to pass all of his subjects above the 30th percentile of those who graduated.

The State Board now permits a student to take these exams any time from age 16 on and if he scores well enough, to leave school if he wishes. Moreover -- and this is perhaps a still more radical step -- the student may have his scores put in permanent records to give employers an idea of his different abilities.

The radical nature of this shift is that we are moving from qualification by process (so many years of schooling) to a qualification by product (certain demonstrated skills in the individual). This shift is, of course, not brand new. The University of Chicago once permitted students to "test out" of many courses and graduate early. And various testing companies have lately offered the College Level Examination (CLEP) and Advanced Placement programs, which have something of the same goal. But most of these programs have been of rather narrow application to high-level students. To offer these programs to all secondary students of a state is something else: A student can appeal over the head of his school to an impartial state agency, and by using objective tests can then skip two years of high school to pursue his own destiny. And the student can use these objective measures as educational or employment credentials throughout his life. That is product, not process.

And we should not expect to stop at the high-school level. Why cannot we expect such test scores to replace the various college degrees? There is no obvious psychological or philosophical principle applicable at a high-school level which becomes suddenly inoperable when we speak of college.

Are We Destroying Tests?

While testing continues to grow, becoming ever more closely tied into our social system, grave abuses must be guarded against. But a remarkable feature of many of our nationalized test programs is the absence of major scandal. Think of it: On the basis of test scores today, students are selected for choice colleges and given large scholarships; they are admitted to graduate school; they are passed through courses which would otherwise cost substantial sums of money and time; they are admitted to professional schools which eventually give them almost automatic membership in a social and financial elite. When we consider the value of these decisions, the clean record of such testing becomes a matter of wonder and pride.

But when we look at individual tests, there is reason for great concern. In one Connecticut district, I am told, when youngsters arrive for their first-grade classification, they may already have practiced 50 "draw-a-man" tests. What does this do to the interpretation of their performance? In another district, there are perhaps 30 parents who meet, before their children have been tested, to go over the items of the Wechsler Intelligence Scale for Children (WISC). All this to make sure their children are on the upper "track"!

Such practices are a great threat to the intelligent use of such individual tests, and the danger suggests a possible remedy: to make the individual test much more like the group standardized test, that is, to regard the items of the test as being simply a small sample from a large, almost infinite, population of such items. One needs to change the items regularly (as in college admissions testing) under great security, but release the old items to the parents and schools. And, in addition, one should make the test, even when orally given, of such a nature that it will not depend on long familiarity with the specific materials. In any case, we need to be sure that test destruction is not a lamentable by-product of well-intentioned legislation, concerning informed consent and several levels of appeal.

What Are Test Scores Worth?

On one general problem of testing, we can be in fair agreement: A great gap exists between the expertise of test development and the amateurish use of the test scores. In fact, the use of test scores, which is most often attacked by opponents, is usually beyond the control of the test producers.

Yet another problem, which has been neglected but to which we might give serious attention, is the worth of the various scores. Consider the question of whether or not one should favor a new program which will cause a two-point rise in math scores but a one-point drop in English scores. Should we go ahead with it? How can we decide? Or consider another program which will take some resources from the slow learners and give them to the gifted. Given these or other trade-offs (and in the real world, trade-offs are always present), how can we select?

A few years ago, some of us tried to look at this question in a way which would tie objective measures together, in a operational system, with top-most educational values. We began by defining a unit of overall educational achievement which we called a bentee, for "benefit T-score" (Page, 1972a, 1973, 1974; Page & Breen, 1973). We found that it would be especially useful to calculate a bentee for any given student, or group, or school, or program. Such a bentee for a program, for instance, would be a weighted sum: the value of each variable, multiplied by its measure for the program and summed across all relevant measures. Suddenly, those questions which seemed hopelessly vague become potentially tractable. But clearly, much work needs to be done on the applications of values to test use, for until we can tame our measures and place them in proper perspective, we will be subjected to strong criticism.

Can We Handle Heritability?

Finally, we come to an issue which seems to underlie many other issues of testing. Our professionals seem to play a kind of hide-and-seek with it: we implicitly acknowledge its importance, yet we turn away from it for a number of reasons. I speak of the controversy over "nature and nurture." How much of test performance is influenced by genes and how much by environment? While it is the touchiest topic of all, let us see if we can articulate some statements which can be assented to, privately at least, by most of us here.

First: We are going to continue using tests, both of ability and of achievement, for many important purposes.

Second: In test scores measuring intelligence, there is a strong influence of heredity. A continuing denial of this point seems simply an error of understanding (like denying regression to the mean from one testing to another) or a simple avoidance of the vast literature on the topic. The broad heritability of intelligence, as measured by our tests, is among the most consensual, and possibly most important, findings in scientific psychology (Erlenmeyer-Kimling & Jarvik, 1963; Scarr-Salapatek, 1971; Page, 1972b). For instance, consider the best-selling text in introductory psychology by Hilgard, Atkinson & Atkinson, 1975. The authors briefly review the literature, showing the classic summary data on the matter, and conclude, "Carrying out the appropriate calculations reveals that above three-fourths of the total variance of IQ scores is associated with genetic factors and one-fourth with environmental factors plus errors of measurement" (p. 417).

Third: Not only tests of intelligence, but also tests of school achievement show substantial heritabilities by use of the classical methods (Loehlin & Nicholas, 1976; Jensen, 1973a). Furthermore, these heritabilities are not simply reflections of an inherited g; there is also evidence that different school achievements, such as English and math may have somewhat different genetic loadings (Martin, 1975).

Fourth: Possible innate group differences, though far cloudier than the points above, underlie much of the current turmoil about testing and its applications. In the 1950's, there was a great push for equalization of the urban and rural worker, and the debate was whether or not there were genetic intellectual differences between the social classes; but now the question is whether or not such differences play any role in any of the test differences between ethnic groups. Great ambiguity exists in the minds of most psychologists and educators, on this question.

Fifth: This ambiguity continues to be a very sore point in wrestling with issues of test fairness, test bias, affirmative action, and many of the current court cases, such as Bakke, "Larry P.," and others. Ethnic differences apparently need to be examined; it does not seem to be enough to look only at income, parental occupation or education, or other measures of social status. This is what was meant by Archibald Cox, defending quotas in medical schools before the Supreme Court. He was quoted as saying, "There is no racially blind method of selection which will draw more than a trickle of minority students" ("Bakke case is argued before Supreme Court," 1977, p. 1). Recent data from the National Longitudinal Study, encountered while we were trying to explain the effects of family size, underscore this point (Page and Grandon, 1977). If

SES is measured in the usual ways, and these ways are constant across ethnic groups, the white lower class would obviously obtain the vast majority of the places awarded to the poor and disadvantaged.

Being aware of this situation is what triggers, again, the need for clarifying the methods of determining heritability across different ethnic groups -- an effort wrestled with by Jensen (1973b) and by Loehlin, Lindzey, and Spuhler (1975), and others. Now may be a time to do our best to cut through the great difficulties of such questions (as posed by Layzer, 1975) and do our best to cast light on the questions (Eaves, Last, Martin & Jinks, 1977).

So now, we begin discussion of these most provocative and central topics in testing. As we said in the beginning, we are arguing about a great deal in the world. Legal implications, minorities, teacher unions, general questions of test use and abuse involve some far-reaching technical, scientific, political, and philosophical issues. I hope we can keep these different aspects properly sorted. In any case, I know we all share the hope that the debate can be conducted, regardless of the degree of variance among us, with the civility and rational tolerance which are the best justification for a university and the best characteristic of any progressive society.

References

Atkinson, R.C. Ingredients for a theory of instruction. *American Psychologist*, 1972, *27*, 921-931.

Bakke case is argued before Supreme Court. *Hartford Courant*, Oct. 13, 1977, pp. 1; 22.

Eaves, L. J., Last, K., Martin, N.G., & Jinks, J. L. A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology*, 1977, *30*, 1-42.

Erlenmeyer-Kimling, L., & Jarvik, L. F. Genetics and intelligence: A review. *Science*, 1963, *142*, 1477-1479.

Houts, P. L. A conversation with Banesh Hoffman. *National Elementary Principal*, July/August, 1975, p. 33.

Jensen, A. R. *Genetics and education*. New York: Harper and Row, 1973a.

Jensen, A. R. *Educability and group differences*. New York: Harper and Row, 1973b.

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. *Race differences in intelligence*. San Francisco: Freeman, 1975.

Loehlin, J. C., & Nichols, R. C. *Heredity, environment, and personality*. Austin: Univ. of Texas Press, 1976.

Martin, N. G. The inheritance of scholastic abilities in a sample of twins: II. Genetical analysis of examination results. *Annals of Human Genetics*, *London*, 1975, *39*, 219-229.

Page, E. B. Seeking a measure of general educational advancement: The bentee. *Journal of Educational Measurement*, 1972a, *9*(1), 33-43.

Page, E. B. Behavior and heredity. (Comment, including Resolution on Scientific Freedom, signed by 50 scientists.) *American Psychologist*, 1972b, *27*(7), 660-661.

Page, E. B. Effects of higher education: Outcome, values, or benefits. In Solmon, L. C., & Taubman, P. (Eds.), *Does college matter*? New York: Academic Press, 1973. Pp. 159-172.

Page, E. B. 'Top-down' trees of educational values. *Educational and Psychological Measurement*, 1974, *34*(3), 573-584.

Page, E. B., & Breen, T. F. III. Educational values for measurement technology: Some theory and data. In Coffman, W. E. (Ed.), Frontiers in educational measurement and information processing. Boston: Houghton-Mifflin, 1974. Ch. 3, pp. 13-30.

Page, E. B. Testing and its enemies. Address at Annual Meeting of the National Council on Measurement in Education and the American Educational Research Association, New York, April 5, 1977.

Page, E. B., & Grandon, G. M. Family configuration and mental ability: Confluence vs. admixture theories. Manuscript submitted for publication, 1977.

Rudman, H. C. The standardized test flap. Phi Delta Kappan, November 1977, pp. 179-186.

Scarr-Salapatek, S. Race, social class, and IQ. Science, 1971, 174 (4016), 1285-1295.

PART I - TESTING AND LEGAL IMPLICATIONS

TESTING: A PARENT'S POINT OF VIEW

Blanche Lewis
Past President
United Parents Association

Though there may be many professional responses to some of the testing issues I will be raising, I want to stress how deeply felt parental concerns are. Parents want the professional testmakers to come down from their ivory towers to where the action is. In this area, as in many others,the professionals must justify testing programs and procedures so that parents are understanding and supportive of their efforts.

One does not need to be professionally engaged in education to be aware of the intense heat that can be generated by the subject of educational testing. Any gathering of parents quickly develops into an animated debate and frequently an anguished discussion of Scholastic Achievement Test (SAT) scores, achievement tests, minimum competency tests, or the whole battery of tests to which children are subjected.

Parents, for the most part, are not opposed to standardized testing; they are, however, seriously disturbed by the use of test results. On the basis of these results, students are labeled, grouped, and sorted. Schools use this information to assign students to classes and to decide how much and what should be taught. From this impersonal and often imprecise information, doors may open or close. Even when test scores are not used for decision making, they follow students thoughout their school experience, influencing the attitudes of parents and teachers and perhaps, even worse, students' expectations for themselves.

The New York City schools' annual testing program, required under the 1970 decentralization law, is given to approximately 700,000 students in grades 2-9. On the high school level, 300,000 students are tested twice a year in reading and mathematics. These test results are used to measure student achievement, but, as mandated by law, they are also used to rank schools and designate so-called "disadvantaged" schools. In addition, results are used by New York City and New York State educational authorities to satisfy requirements imposed by federal law in determining the school system's eligibility for federal funds.

It seems as though a whole new set of purposes have been legislated for tests. Parents are asking, at what price to the students. As long ago as 1960, the United Parents Association (UPA) conducted a testing study which, if published today, would still be relevant. Then, as now, parents felt that tests could be instruments for achieving better education if they were followed up by effective research. In other words "test the testing." UPA cautioned against exclusive dependence on standardized tests for categorizing students. Tests are only one tool that should be used in guiding students. However, "in spite of its weaknesses," said UPA "testing is necessary, valuable and indispensable if the results are properly interpreted."

UPA recommended that Boards of Education set policies requiring principals to present and interpret test results in a comprehensible way to parents. Because of teachers' limited knowledge about testing, UPA also recommended that in-service courses in testing be required before teachers are permitted to administer and interpret standardized tests. Although the testing of non-English speaking children and "culturally deprived" children had not been included within the scope of its study, UPA did recommend that the development of appropriate tests be intensified and accelerated.

Significantly, at the same time some parents were pursuing these goals through discussion and negotiation, others were turning to the courts. A suit, brought by a Long Island parent in 1960, to compel the district school board to show him his child's test scores and records resulted in a court clarification of State Education Department policy "that parents have the right to inspect their children's school records and that the schools must study carefully and plan for student records and for their proper interpretation to parents."

Although parents are now entitled to receive such information, there appears to be inadequate teacher training programs designed to help in communicating and interpreting test results to parents. And in addition, parents are often shunted aside with the argument that as laymen they simply cannot understand the testing field.

Another problem is mechanized scoring of tests. Computer technology and teacher contracts have had a marked effect on the use of tests as a diagnostic tool. In New York City, as in other cities, teachers no longer hand score tests. Even though computers have the capability to give us an unlimited amount of information, generally only the results of the standardized tests are returned to the teacher. Teachers do not receive information which would enable them to diagnose individual learning problems to help improve students' skills.

All too often, testing, as perceived by parents, is not so much a helpful tool but a hindering one which excludes students from the best education they deserve. Putting students into groups on the basis of their test scores becomes a double-edged sword. In many instances, the aim is to group students according to their abilities, challenging gifted students and offering those less able special help and training. Unfortunately, using test scores this way contributes to a lessening of opportunity and expectation for many students.

What can parents do about all of this? In our frustration shall we demand an end to all standardized testing? Some parents, principals, teachers and school boards have decided to do so. Others continue to believe that students need to be tested, giving the results to the community. The drop in national achievement scores, along with mounting concerns about the cost of education, has supported this view and led to calls for greater accountability and improvement in the quality of education. On the other hand, resistance to demands for reform have led parents to seek new avenues for solutions, and for many, the courts represent such an avenue. Since 1954, the courts have increasingly challenged educational decisions that have not coincided with society's goals for the schools.

The influence of federal courts in desegregating schools is widely known. The courts have also been extremely active in other arenas: the rights of teachers, the rights of children, and state financing of public education. The courts have certainly not been insensitive to concern for the quality of schooling. For example, in Brown v. Board of Education (1954), the Supreme Court determined that the quality of every child's education is poorer when schools are racially segregated. In Serrana v. Priest (1971), the California Supreme Court determined that financial resources must be equalized because the quality of education depends upon resources.

Recently, considerable controversy has arisen regarding the role of both federal and lower courts in local educational decision making. Traditionally, the courts have tended not to intervene in school board affairs unless the constitutional rights of an individual are abridged or unless the board has failed to act when it had the responsibility to do so. Educational policy has historically been the function of legislatures and school boards. In recent years, there have been indications that the era of educational immunity from legitimate checks has come to an end. The federal government, through the passage of the Elementary and Secondary Education Act of 1965, has provided the states with a limited means for checking schools.

Many see the movement toward establishing programs to determine minimum competencies as another check in assessing school performance. Here too, though some of the language is new, many of the problems remain the same. Parents fear that once again the onus of measuring the school's performance falls on the child. As standards are set, parents fear the minimum may become the maximum. The question of how to force the educator to educate remains unsolved. As a result, parents look more to the courts for solution.

The issue of the role of the courts in local school policy making came to national prominence with the landmark Hobson v. Hanson decision (1967). Justice J. Shelly Wright ordered the desegregation of the 90 percent black *de facto* segregated Washington D.C. school system. He also required Washington to equalize resources available to all schools. In Hobson v. Hanson the courts abolished ability grouping and challenged the use of standardized tests as a criterion for such grouping of disadvantaged children.

In the Aspira v. Board of Education Consent Decree, (August, 1974), the Board of Education agreed to furnish programs "to those children whose English language deficiency prevents their effective participation in the learning process and who can more effectively participate in Spanish to develop language, mathematics, social studies and science skills." The decree also mandated for classification purposes that improved assessment instruments be developed to measure reading, writing, speaking and understanding in English and in Spanish.

Last year the New York City testing program was halted by the State Supreme Court and by the Board of Education, after it became known that in some schools 3000 children had been drilled on forthcoming test questions. Although the controversy focused on the secure and non-secure labels used to identify the tests, a wide mixture of other issues were also highlighted. These included the professional pride of teachers and administrators, the increasing demand of parents for improved school performance, and the politics of education. The courts ultimately ruled that the original injunction was an unlawful interference in educational policy judgment, but also said "this is not to say that there may never be gross violations of defined public policy which the courts would be obliged to recognize and correct."

Another court case involved a 5 million dollar "educational malpractice" suit brought by a young Lindenhurst, Long Island man who said his school district had allowed him to graduate from high school although he was functionally illiterate. Here too, the case was dismissed. However, Justice Paul J. Baisley recommended review by a higher court because of the vital issue raised: whether or not a school can be held accountable for a student's illiteracy.

Given this sort of legal precedent and given the degree of dissatisfaction with many school systems, it is likely parents will be suing the schools to enforce what they see as a right of their children to quality education. Most parents see testing as a necessary part of quality education. Yet many parents distrust testing because of the way they believe it is being used. Parents will not cease in their determination to secure more effective testing and a better education for their children.

MUST EDUCATORS BE LAWYERS TOO?

Anthony J. Polemeni, Director
Office of Educational Evaluation
New York City Schools

Today many educational matters, such as school funding, integration, and accountability, are being decided not by educators -- whom some might consider best qualified -- but by lawyers, judges, and juries. Perhaps this is part of the syndrome, characteristic of our time, in which all matters from womb to tomb are decided on the basis of law rather than ethics. The force of this argument is well evidenced by two legal cases concerned with testing programs in New York City: the imposition by the United States Department of Health, Education and Welfare (HEW) of minority norms on an experimental testing program and a ruling by the New York State Supreme Court concerning the validity of a standardized test administered citywide in New York City.

The crux of the first case was either a local school district within New York City would administer a minority-normed test adapted by HEW or HEW would withhold $1,000,000 in federal funds from the district. The matter developed as follows: In February 1973, a regulation published by HEW to guide applicants for Emergency School Aid Act (ESAA) funds stated, in part, that applicants must cooperate in program evaluation, including the administration of achievement tests. These tests, along with other instruments, were to determine the effectiveness of ESAA programs in improving race relations, reducing minority group isolation, and improving basic skills. HEW, however, through its contractors, determined that existing nationally standardized tests were invalid for, and biased against, minority group students and must be renormed.

While these experts were determining what tests were to be restandardized and what sample was necessary for restandardization, HEW personnel were advising local educational agencies of ESAA provisions and providing them with application forms. Shortly thereafter, a letter was sent to local educational agencies asking for their cooperation in a norms testing effort scheduled for May, 1973. The program was explained in detail, and the letter ended with this sentence: "The participation of your district in these testing activities has no bearing on ESAA grant procedures or decisions."

However, when the experimental testing program began, charges were made by personnel both in the district and at the Board of Education that the special norms for minority children would result in lower standards and that these students were not as capable of learning as other students. On the basis of this charge, one of the local school districts in New York City refused to allow HEW to administer the tests. As a consequence, the district was notified that its request for ESAA funding was disapproved -- in spite of HEW assurance that its approval was not contingent upon participation in the experimental testing program. At this point, the Board of Education entered a Civil Action against HEW to have the $1,000,000 grant reassigned to the local school district.

In my deposition for this case, I said the following:

> This program involves the use of a test known as the California Achievement Test, which is a standard

> commercial device intended to measure the educational achievement levels of elementary school children. It is intended by the defendants in this pending action to give this test to a selected group of third, fourth and fifth grade elementary school students in District 19 for the alleged purpose of developing dual standards through a renorming process employing a commercially developed test.
>
> In my view, this constitutes an experimental program which has questionable validity in needs assessment evaluation and which is beyond the intent of the evaluation contemplated for the programs.
>
> The basic substantive objection of the plaintiffs related to the artificiality of the scoring norms established by way of the testing procedure. It is a procedure which can only be described as tailored testing and which permits an examiner to achieve any result which he may desire. The ultimate result of such manipulation can only be an invalid, prejudicial and incorrect image in relation to the educational achievements of the minority students intended to be tested. It would in effect create a Federal Policy which mandates second-class citizenship for all minority students.

Ms. Elaine B. Landrum, Superintendent of Community School District 19, in her deposition in this matter said the following:

> As an expression of personal opinion, representative of the thinking of virtually all of the parents and school children in Community School District 19, I have no hesitation in stating that there is a universal feeling of abhorrence at the concept of two separate norms which can only have the ultimate effect of seriously distorting their educational achievement level when confronted with the need to pursue their education in high school and college. The handicap imposed upon these children by reason of the inaccurate self image that is bound to result is so serious that it requires no further elaboration.
>
> Even more serious is the impact that may be expected when these children, while laboring under a false impression as to their educational achievement, are confronted with the fact that they have been treated differently.

> The net effect could well be that in the future, minority students may be denied equal access to educational facilities as a consequence of the "suspect" nature of their educational achievement records.
>
> One of the stated purposes of the basic legislation is the elimination of minority group isolation. If the procedure contemplated here is permitted to be implemented, the actual and inevitable result will be to increase minority group isolation rather than reduce it.

These depositions, along with many other documents and the good will of all parties concerned, resulted in a Stipulation of Agreement and Discontinuance dated February, 1974. The terms of this agreement assured the Board of Education that no separate norms would be developed for minority and/or non-minority children as part of, or as a result of, the ESAA evaluation process. Upon reaching this resolution, the entire legal action was discontinued.

The second case in which the courts became involved in a New York City testing program dealt with a determination of whether or not a standardized test to be administered citywide in grades 2 through 9 was valid. Essentially, the case centered around the following:

Each year during the last week in March or the first week in April, a nationally standardized reading test is administered to some 700,000 New York City public school students in grades 2 through 9. The purpose of the test is to provide data for policy and promotional decisions, as well as to satisfy a mandate of the New York State Legislature that all schools be ranked according to the percentage of pupils reading at and above grade level. The test to be used is determined through competitive bidding procedures and, for school year 1976-77, the selection was the CTB/McGraw Hill Comprehensive Test of Basic Skills.

The citywide test was scheduled for administration on March 29, 1977, but, several weeks prior to that date, copies of the test were circulated as "Practice Test Materials" in one of the thirty-two local school districts of New York City. Although it was subsequently determined that the test was actually administered to only 34 classes in the district, the story of the inadvertency was picked up by the press, radio, and TV, and an immediate reaction resulted in some quarters that those students who had been exposed to the test would have an unfair advantage over those who had not.

The matter was taken to court and, on April 11, 1977, a ruling by the State Supreme Court said in part, "To administer these examinations now would be a futile exercise. To rely upon the results would be unfair to those who have lacked prior familiarity with the examinations. ...The only appropriate conclusion, in this Court's view, is for the comprised examinations to be entirely discarded, and new examinations, suitably secure, to be administered." This ruling by the State Supreme Court came in the face of affirmation by the Chancellor of New York City schools that the examinations had, in fact, not been compromised to such an extent as to render them invalid.

The lines, therefore, were cleanly drawn: either the Chancellor, New York City's leading educator, or the State Supreme Court was going to determine whether the Citywide Testing Program was or was not valid. When the State Supreme Court ruled that the examinations were invalid, the matter was taken to the New York State Court of Appeals where the decision was reversed. That portion of the Court of Appeal's decision which will be of interest to educators throughout the United States reads as follows:

> Whether the courts have power to enjoin, even temporarily, the administration of examinations to school pupils based on contentions that the integrity of the examinations had been fatally compromised is the issue presented for our consideration. Special Term, affirmed by the Appellate Division, granted a preliminary injunction against the scheduled administration of comprehensive reading and mathematics examinations to pupils throughout the entire New York City school system. There should be a reversal. The temporary injunction issued by the courts below was an unlawful interference with an educational policy judgment made by the appropriate school authorities in exercise of constitutional and statutory power. Whether or not an examination had been so compromised as to strip it of validity as a device for measuring educational achievement is a matter committed to the professional judgment and discretion of those responsible for the administration of the public schools. It is not a matter for the courts.

In the final analysis, New York City's Chancellor of schools was permitted to determine whether a test was or was not valid, and HEW was prohibited from administering minority-normed tests in New York City.

But we had better be realistic about it. These two cases are only the beginning of the battle. Our authority and jurisdiction in educational matters is being threatened anew each day by politicians, local, state and federal agencies, and the courts. The question is, are the schools to be run by educators or by jurists? If the answer is the latter, then educators must become lawyers too.

LEGAL IMPLICATIONS OF INTELLIGENCE AND COMPETENCY TESTING

Merle Steven McClung, Esq.
Staff Attorney
Center for Law and Education
Harvard University

The abundant literature detailing the limitations of standardized test results does not appear to have curtailed reliance upon them in making determinations about intelligence and competence. Since these determinations have profound effects upon a person's educational and occupational future, serious legal questions are often raised.[1]

While intelligence and competency testing have many differences, a common legal focus is likely to be whether the particular test or test practice discriminates against a person or persons on account of race, color, national origin, or other factors, such as sex or handicap. Laws relating to privacy, individual choice, and general fairness may also pertain. In applying these legal standards, courts often rely upon the various professional standards relating to testing. This paper will discuss some of these legal and professional standards as applied to two testing situations: (1) IQ testing to determine which children should be placed in special classes for the mentally retarded, and (2) competency testing to determine which students will receive a high school diploma. The primary example of the first situation will be the Larry P. v. Riles case currently being tried in California; and the primary example of the second will be Florida's "Educational Accountability Act of 1976."

I. Intelligence Testing and Mental Retardation

Cultural bias is the charge most frequently advanced by minorities against standardized intelligence tests. Perhaps the most obvious example of cultural bias and consequent misclassification is when a child is given an IQ test in a language other than his or her own. Placements, reflecting this kind of linguistic discrimination, in educable mentally retarded (EMR) classes based upon low IQ scores were successfully challenged in Diana v. California State Board of Education,[2] a consent agreement which, among other things, precluded English-only tests for children whose primary home language is other than English.[3]

Current controversy, however, centers on the more subtle kind of cultural bias which is being challenged in a California case titled Larry P. v. Riles.[4] Plaintiffs in that case argued that the stigmatizing EMR label was applied on the basis of IQ tests which penalize unfamiliarity with white middle-class values to a disproportionate number of black children. Defendant school officials argued that the disproportionate number of black children in EMR classes reflected poor infant care, inadequate nutrition, and other environmental factors which may create "a poor genetic pool for all races ghettoized in the inner city."[5] The Court initially concluded that the defendants had not sustained their burden under the equal protection clause of proving a rational relationship between scoring on the questioned IQ test and the ability of black students to learn, and issued, therefore, a preliminary injunction against future IQ testing and future reevaluations.[6]

The trial on the merits in Larry P. is now being held. Similar cases have been brought in other jurisdictions,[7] and some of them seek monetary damages for alleged educational malpractice in misclassifying children for special education programs.[8]

Legal Standards

The equal protection challenge against discriminatory testing is more difficult today as a result of the recent ruling of the U.S. Supreme Court in Washington v. Davis.[9] In that case, which involved a police personnel test used in the District of Columbia, the Court ruled that the disproportionate racial impact of a test is not sufficient to establish an unconstitutional racial classification without proof that it reflects a racially discriminatory purpose.

As a result of Washington v. Davis, plaintiffs in testing cases are more likely to emphasize Title VI of the Civil Rights Act of 1964 which incorporates an effect (rather than purpose) standard, prohibiting practices which have the effect of discriminating against individuals on the ground of race, color, or national origin.[10] Thus plaintiffs in Larry P. have amended their complaint to include a Title VI claim.

Application of these Title VI standards to public school testing programs is indicated by an H.E.W. memorandum stating that there may be a Title VI (or Title IX, sex discrimination[11]) violation where the school has failed "...to adopt and implement procedures to insure that test materials and other assessment devices used to identify, classify, and place exceptional children are selected and administered in a manner which is non-discriminatory in its impact on children of any race, color, national origin, or sex...."[12] (emphasis added).

Adaptive Behavior

While an important issue, cultural bias should not be the sole or even the primary focus in determining whether or not a child should be classified as mentally retarded and/or treated like a mentally retarded person.[13] Removing culturally biased test items will not insure proper classification because other requirements must be met to justify the label of mentally retarded. The American Association on Mental Deficiency (AAMD) defines mental retardation as "significantly sub-average general intellectual functioning existing concurrently with deficits in adaptive behavior, and manifested during the developmental period."[14] (emphasis added).

Significantly sub-average intellectual functioning is defined as two or more standard deviations from the mean or average of the test; that is, a score of 68 on the Standford-Binet Intelligence Scale and 70 on the Wechsler Intelligence Scale for Children.

Adaptive behavior, is defined as "the effectiveness or degree with which the individual meets the standards of personal independence and social responsibility expected of his age and cultural group."[15] Measurements of adaptive behavior are necessary to eliminate "the six-hour retarded child" who is considered normal by family, friends, and community, but labeled retarded by the school.

The concept of adaptive behavior is important in student classification because it accounts for an aspect of intelligence not measured by standardized IQ tests. IQ tests measure a narrow dimension of human intelligence. Some people call this "school" intelligence as opposed to the kind of "life-skill" intelligence measured by adaptive behavior. The AAMD offers a set of adaptive behavior scales. Other groups also have developed or are developing similar "objective" measures of adaptive behavior.[16]

These dual criteria--IQ and adaptive behavior measures--are incorporated in the consent agreement in Lebanks v. Spears,[17] a case challenging the placement of children in classes for the mentally retarded in Louisiana's Orleans Parish public schools. Lebanks adds a third prerequisite to the mentally retarded label; the child must still be rated substantially sub-normal on both measures after the effects of sociocultural background have been taken into account. These criteria are also incorporated in a consent agreement in The Rhode Island Society for Autistic Children (RISAC) v. Board of Regents of Education of Rhode Island.[18] The RISAC agreement further provides: "A parent may waive this requirement, but only after having been informed in writing that the American Association on Mental Deficiency recommends that no child with an IQ similar to that of his child be labeled retarded."[19]

New Federal Standards

New federal standards designed to protect children in evaluation procedures relating to special education are incorporated in the Education for All Handicapped Children Act of 1975, Section 504 of the Rehabilitation Act of 1973, and their implementing regulations.[20] Any child who is thought to be handicapped must be fully evaluated before being placed in either a regular or special program.[21] These standards also provide, inter alia, that the evaluation procedures and materials used must be selected and administered so as not to be racially or culturally discriminatory,[22] must be administered in the child's native language or primary mode of communication unless it is clearly not feasible to do so,[23] and must be validated for the specific purpose for which they are used and administered by trained personnel.[24] Placement decisions cannot be based on any one single evaluation procedure, but must instead be based on a variety of factors, including well documented information concerning aptitude, achievement, teacher recommendation, physical condition, socio-cultural background, and adaptive behavior.[25]

II. Competency Testing

The State of Florida is in the forefront of a competency testing movement which is sweeping public education in this country. Pursuant to Florida's "Educational Accountability Act of 1976,"[26] students in that state who do not pass a functional literacy test before they graduate will be given a certificate of attendance rather than a high school diploma. On the surface, the new law seems fair enough, but there are problems. The advance notice, the phase-in period, and the match between instruction and the test are all alleged to be inadequate, and a large number of students are expected to fail the test. A disproportionate number of these students are likely to be black and Hispanic students who have suffered from inferior educational opportunities. As a result, lawyers are considering legal action against the proposed use of this test.

This emphasis on competency testing is, of course, a response to the widespread public dissatisfaction with the measurable outcomes of public schooling. A number of studies[27] indicate that, whatever definition of literacy is used, substantial numbers of Americans are not literate. The response of Florida and other states to functional illiteracy, unfortunately, has been competency testing programs which unfairly penalize some students.

Because the minimal competency requirement as a prerequisite to a high school diploma is a new development in most states, to identify the strongest legal arguments for or against it and to predict the judicial response is difficult. Legal challenges to competency testing programs will have to be tailored to specific factual situations, but most will probably emphasize the following: (1) racial discrimination, (2) inadequate notice and phase-in periods, and (3) inadequate match between test and instruction.[28]

Racial Discrimination

A disproportionate number of the students who do not pass competency tests are likely to be black and Hispanic students who have suffered from inferior educational opportunities. For example, the Palm Beach Post-Times reported the results of a competency test administered in the Palm Beach County Public Schools: "A massive disparity between the performance of whites and nonwhites reflected in a trial run of the test last year was repeated with 42 percent or 576 minority children failing as opposed to 8 percent or 282 white children."[29]

As mentioned in the discussion of intelligence testing, the racially disproportionate effect of a test is sufficient under the Title VI standard to shift the burden of justification to the defendants.[30] Where competency testing programs have a disproportionate effect on Hispanic children, courts will also be asked to consider Lau v. Nichols[31] and related HEW memoranda[32] requiring public schools to take affirmative steps to remedy the linguistic exclusion of non-English speaking children from educational opportunities.

The point is not that minorities should be excluded from meeting the same standards of competence as whites, but rather that programs must be developed that do not perpetuate the effects of past discrimination. Indeed minorities who, in disproportionate numbers, have not been properly instructed in the basic skills are in the forefront of those calling for increased attention to these skills.

The case for illegal discrimination against minority students will be stronger where the disproportionate racial effect of competency testing is combined with other evidence of unfairness, such as inadequate notice and phase-in periods and inadequate match between the test and instruction. These problems, which can also harm students regardless of race, and other legal approaches, which are not dependent upon racial effect, are discussed below.

Inadequate Notice and Phase-In Periods

Imposing the competency testing requirement on a students one year before graduation, as in Florida, means that most of these students will have spent their first ten or eleven years in the school system without notice or knowledge that passing a competency test would be a condition to acquiring a diploma. It is likely that many, if not most, of those students failing the test would have studied (and teachers taught) differently in earlier years had they been given such notice.

Traditional notions of due process should require adequate prior notice of any rule which could cause irreparable harm to a person's educational or occupational prospects.[33] Whatever notice is considered fair (first grade? fourth grade?), notice after most of one's educational program is already completed seems clearly inadequate.

Inadequate Match Between Test and Instruction

The test instruments used in competency testing will probably be subject to the validation standards developed by the American Psychological Association (hereafter "APA Standards").[34] Courts have relied upon the APA Standards in other testing cases.[35] In addition, competency tests should have "curricular" and "instructional validity," as defined below.

Most people would agree that fairness requires that a school's curriculum and instruction be matched in some way with whatever is later measured by the test. The test would be unfair if it measured what the school never taught. This concept is not included in the APA Standards[36] but is of primary importance in considering the validity of any competency test.

Curricular validity is a measure of how well test items represent the objectives of the curriculum. An analysis of curricular validity would require comparison of the test objectives with the school's course objectives.

A competency test should also have what may be called *instructional validity*. Even if the curricular objectives of the school correspond with the competency test objectives, there must be some measure of whether or not the school district's stated objectives were translated into topics actually taught in the district's classrooms. Instructional validity obviously does not require prior exposure of the student to the exact questions asked on the test, but it does require actual exposure of students to the kind of knowledge and skills which would enable a student to answer the test questions. This will present difficult proof problems in some cases; in others, it will be relatively easy to show that the test is measuring what the school never taught.

A competency test measuring life-skills and used as a basis for denying a diploma, when such skills were never taught in the school, is arguably so arbitrary as to violate due process of law.[37] A competency test lacking curricular or instructional validity may also violate substantive due process because the school, rather than the student, can be faulted for poor performance on the test.[38]

There is evidence that the competency test used in Florida's Palm Beach County schools lacks both curricular and instructional validity. One Area Superintendent stated that basic skills were not stressed in the schools during ten years of "curriculum upheaval."[39] And a Director of Instructional Services stated that the test expects some knowledge--like how do you apply for a job or how do you use an extension service--that has not been taught in the classrooms.[40] Perhaps many schools could show that their instructional programs have, in fact, been geared to the basic literacy skills measured by their versions of the competency test, but this may not be true of many "adult performance level" tests which measure content knowledge in addition to literacy and the ability to apply basic literacy skills to real-life situations.[41]

Conclusion

We presently use standardized test results to decide, among other things, which children will be placed in special education classes, which students will receive a high school diploma, which high school graduates will go to college, which college graduates will enter a profession, and even which citizens will secure employment. The basic questions center on whether this use of standardized test results is fair and consistent with the principles of our constitutional democracy. The tests and testing practices currently in use indicate that we fall far short of these goals; and as a result, the courts will continue to play an important role in shaping the parameters of testing.

FOOTNOTES

1. This paper is a summary of an article by the author titled "Testing for Intelligence and Competence: Legal Implications," available from the Center for Law and Education, Gutman Library, 6 Appian Way, Cambridge, MA 02138.

2. C.A. C-70-37 R.F.P. (N.D. Cal. 1970).

3. For a discussion of the problems of using translated IQ tests with Hispanic children, see A. Padilla and B. Garza, "IQ Tests: A Case of Cultural Myopia," in P. Houts (ed.), *The Myth of Measurability* (1977) at 124-36.

See generally J. Morris, "What Tests Do Schools Use with Spanish-Speaking Students?" 15 *Integrated Education* 21 (March-April, 1977).

4. *P*. v. *Riles*, 343 F. Supp. 1306 (N.D. Cal. 1972), *affirmed* 502 F.2d 963 (9 Cir. 1974).

5. Education Daily, Oct. 18, 1977, at 3. See also *P*. v. *Riles*, 343 F.Supp. 1306, 1310-11 (N.D. Cal. 1972).

6. This preliminary injunction was initially limited to black children served by the San Francisco Unified School District, but was later extended to protect all black children in California.

7. *Lebanks* v. *Spears*, *infra* note 17; *RISAC*, *infra* note 18.

8. For legal citations and summary of cases, see "Testing For Intelligence and Competence," *supra* note 1.

9. 96 S.Ct. 2040 (1976).

10. 42 U.S.C. §2000d (1974). For regulations implementing Title VI, see 45 *C.F.R*. Part 80 *et seq*., especially Parts 80.3 and 80.5

11. Title IX of the Education Amendments of 1972 prohibits discrimination on the basis of sex in education programs or activities benefiting from Federal financial assistance, 20 U.S.C. §1681(a) (1974). For H.E.W. regulations on Title IX, see 45 C.F.R. Part 86.1 *et seq*.

12. H.E.W. Memorandum by O.C.R. Director, M. Gerry, "Identification of Discrimination in the Assignment of Children to Special Education" (August 1975). This memorandum also indicates other procedures to be used in evaluating children for special class placement, including adaptive behavior measures.

13. Some school systems have been moving away from labels such as mentally retarded since they are considered stigmatizing and unduly categorical, but most of the problems persist if the child is treated as if one of the former labels applied.

14. H. Grossman (ed.), Manual on Terminology and Classification in Mental Retardation (published by the American Association on Mental Deficiency, 1973 ed.) at 11.

15. *Id*.

16. Adaptive behavior scales have also been developed by Jane Mercer. For description of her work on adaptive behavior, see J. Mercer, Labeling the Mentally Retarded (1973).

17. C.A. No. 71-2897 (E.D. La., April 1973).

18. C.A. No. 5081 (D.R.I., Stipulations on Settlement, September, 1975).

19. *Id*.

20. The Education for All Handicapped Children Act of 1975 is codified at 42 U.S.C. §1401 *et seq*., and its implementing regulations at 45 C.F.R. Part 121a.

Section 504 of the Rehabiliation Act of 1973 is codified at 29 U.S.C. §794, and its implementing regulations at 45 C.F.R. Parts 81 and 84.

21. 45 C.F.R. §84.35(a).

22. 20 U.S.C. §1412(5) (c) and 45 C.F.R. §121a.530(b).

23. 20 U.S.C. §1412(5) (C) and 45 C.F.R. §121a.530(a)(1); see also 45 C.F.R. §84.35(b) (3).

24. 45 C.F.R. §84.35(b) (1) and 45 C.F.R. §121a.532(a) (2) & (3).

25. 20 U.S.C. §1412(5) (c) and 45 C.F.R. §121a.533(a) (1); see also 45 C.F.R. §84.35(c).

26. Florida School Laws (1976 ed.) Ch. 299.55 *et seq*., especially ch. 229.814, 230.2311 and 232.24.

27. See, for example, Public Health Service, Literacy Among Youths 12-17 Years (December 1973), HEW Pub. No. (HRA) 74-1613. This study concluded that an estimated one million American youths 12-17 years old probably could not read as well as the average fourth grader, and thus could be called illiterate.

28. For a more detailed legal and educational analysis of these issues, see M. McClung, "Competency Testing: Potential for Discrimination," 11 Clearinghouse Review 439 (September 1977) (hereafter cited as "Competency Testing").

29. Palm Beach Post-Times, May 10, 1977.

30. Blacks and other minorities may not have to prove discriminatory intent to establish an equal protection violation in school districts subject to prior discrimination. In comparable situations the federal courts have held that practices which carry forward the effects of prior discrimination are unconstitutional. See, for example, Gaston County, N.C. v. United States, 89 S. Ct. 1720 (1969) (voting rights); and McNeal v. Tate County School Dist., 508 F.2d 1017 (5 Cir. 1975) (ability grouping).

31. 94 S.Ct. 786 (1974).

32. See HEW Memorandum by OCR Director J. Pottinger, "Identification of Discrimination and Denial of Services on the Basis of National Origin," (May 25, 1970). See also subsequent memoranda, "Evaluation of Voluntary Compliance Plans..." (Summer 1975), and "Application of Law Remedies" (April 8, 1976).

A claim of linguistic discrimination could also be based upon 20 U.S.C. §1703(f) which provides that "No state shall deny equal educational opportunity to an individual on account of his or her race, color, sex, or national origin, by...(f) the failure by an educational agency to take appropriate action to overcome language barriers that impede equal participation by its students in its instructional programs."

33. For legal authority and analysis, see "Competency Testing," supra note 28, at 444.

34. American Psychological Association (APA), Standards for Educational and Psychological Tests (1974). This publication was prepared by a joint committee of the APA, the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME).

35. For legal authority and analysis, see cases cited in "Competency Testing," supra note 28, at footnote 59.

36. The APA limits its discussion of validity standards to consideration of the concepts of criterion-related validities, content validity and construct validity. Most probably, this limitation occurs because the standards are promulgated primarily to regulate the conduct of organizations which develop and market tests and organizations which purchase those tests. Also, the standards are written by test designers and institutional users.

These two factors may explain why the APA Standards do not include some considerations such as curricular and instructional validity which are important to test takers. Content validity is a related, but distinguishable, concept.

37. See *supra* note 33.

38. In an analogous case, *St. Ann* v. *Palisi*, 495 F.2d 423, 426 (5 Cir. 1974), the Fifth Circuit held that a school board regulation violated substantive due process because it allowed school children to be suspended for their parent's misconduct. For summary, see "Competency Testing," *supra* note 28, at 447.

39. *Palm-Beach Post Times*, April 3, 1977.

40. *Miami Herald*, July 21, 1977.

41. See generally W. Spady, "Competency Based Education: A Bandwagon in Search of a Definition," 6 *Ed. Researcher* 9-14 (January 1977); G. Madaus and P. Airasian, "Issues in Evaluating Student Outcomes in Competency Based Graduation Programs," 10 *J. of Res. & Dev.* in Ed., 79-91 (Spring 1977).

PART II - TESTING OF MINORITIES

ALTERNATIVES TO QUOTAS

Richard B. Darlington
Associate Professor of Psychology
Cornell University

Ten or fifteen years ago, many believed that affirmative action programs would not be needed to substantially increase the number of minority students in our colleges and universities. It was suggested that standard admissions tests were biased against blacks and other cultural minorities, and that if this bias could be eliminated, the number of black students in our colleges and professional schools would increase naturally, without affirmative action.

It is now known that the tests produced by the Educational Testing Service (ETS) and the American College Testing Program (ACT) are not substantially -- or even noticeably -- biased against black students or other minorities in relation to grade-point averages in colleges. The relevant data were summarized best by Linn (1973), though I have seen a number of individual studies since then which support the same general conclusion.

Before proceeding further, let me distinguish between a culture-fair selection procedure and a culture-fair test. A selection procedure is a set of rules which determines, in the last analysis, which applicants are admitted to an institution. A test, on the other hand, is a measure offered to the designers of selection procedures, and is typically one element of a selection procedure.

In my opinion, a culture-fair test for college admissions purposes is one that correlates with cultural variables (for instance, race) no more highly than the criterion variable does. The criterion variable is what the test is supposed to predict, an idea proposed in 1971 by Robert Thorndike. The Linn data show that the ETS and ACT tests are approximately culture-fair by this definition, if college grade-point averages are used as the criterion variable. A culture-fair selection procedure is a little different from that of a culture-fair test. In my opinion, a selection procedure is culture-fair if it does the best possible job in selecting people who will succeed at the task for which they are selected. If we imagine a selection procedure which consists entirely of selecting the people with the highest scores on the ETS or ACT tests, then the Linn data summary shows that this selection procedure would not be biased against blacks, provided the college grade-point average is used as the criterion variable. In other words, eliminating bias in our tests will not increase the number of blacks and other minorities admitted to our selective colleges and professional schools. Therefore, only affirmative action will effect this increase.

ETS recently released a study on the importance of affirmative action in increasing the number of minority students now admitted to law school who would not have been admitted if affirmative action programs were abolished. In this study, Franklin Evans asked the ETS researcher what would happen if the Supreme Court ruled in favor of Bakke, and schools were forced to abandon affirmative action programs, accepting minority students at the same rate as majority students with similar undergraduate grade-point averages and similar scores on the Law School

Admissions Test. It should be noted that, under existing conditions, approximately 1700 black applicants were admitted to one or more of the nation's 163 accredited law schools in 1975-1976. If affirmative action had not existed, that number would be cut from about 1700 to about 700. Thus, if affirmative action were abolished, about 58% of the black students currently being admitted to law school would no longer be admitted. Even worse, the number of black applicants admitted to the ten most selective law schools would be only ten, an average of one black student per school.

This analysis shows what would happen if admissions were based on culture-fair tests. If admissions were based on culture-fair selection procedures, as I have defined the term, even fewer black students would be admitted. Thus, the abandonment of affirmative action programs would result in major changes from current practice.

Having established that affirmative action has an important effect, let me state some arguments for affirmative action.

First, grade-point average within the selecting institution is not the ultimate criterion. It is reasonable to believe that minority lawyers or physicians or other professionals will provide more services to our minority communities than non-minority professionals will. In that sense, minority applicants are more qualified than majority applicants.

In terms of the Bakke case, an important aspect of this argument is the following: To deny affirmative action, the Supreme Court would essentially have to deny the right of selecting institutions to make any judgments about what kinds of skills are needed in the real world as opposed to school. A ruling in favor of Bakke would mean that if a nationwide shortage of eye surgeons were foreseeable, it would be unconstitutional for medical schools to give preference to students expressing a special interest in eye surgery, since it would also be unconstitutional to give preference to students most likely to meet the need for doctors in minority communities.

The point I have just made has been made earlier, and more eloquently, by others. However, in my opinion, a second argument for affirmative action has not been made before: That each selecting institution might view itself as an agent of society as a whole. Thus, each medical school should not solely attempt to select the people who would make the best doctors, but rather select people who will contribute more to society as doctors than they will contribute in some other profession.

This argument is particularly relevant to the Bakke case, since Bakke, is already a trained and, presumably, productive engineer. Even if he paid his own tuition to engineering school, his education as an engineer was subsidized by society as a whole because tuition does not normally cover the cost of a college education. He is now asking to be subsidized for a second round of professional training, even if it means denying some other applicant just entering a professional school.

A third argument for affirmative action is that we still do not know entirely what is actually being measured by so-called aptitude tests. I recently came across a remarkable study published in 1923, which has somehow been ignored for decades by writers on minority testing. The study examined the Stanford-Binet IQ scores of children raised on canal boats in England around 1920. Because these children lived on constantly moving boats, they did not attend school. One study found that 532 canal-boat children had attended an average of only 17 days of school in the previous year. In other ways, the children lived normal and healthy lives. Quoting from the report,

> The Canal Board population as a rule, is born, lives, and dies on the boats. It is stated that at least 4 pounds a week is earned on an average, even in bad times; at other times considerably more. Some of the men even own their own boats...When the boats remain in a town for loading and unloading, the children ...attend places of entertainment, such as cinemas, etc., and have money to spend. Many of the children are well dressed, clean, and appear fairly intelligent...The majority were found to be anxious to talk...To sum up, these children in respect to health, cleanliness, morality, feeding, etc., are fully equal, if not superior, to town dwellers of a similar character...they are not mentally defective, as is generally understood by that term. (Gordon, pp. 229-231)

The Stanford-Binet IQ scores of these children were about normal at ages 5 and 6, but decreased rapidly with increasing age. The children at ages 12 to 14 had an average IQ score of only 56 -- a level considered severely retarded. Yet, as we have seen, they did not act retarded at all -- they simply did not go to school. It is remarkable that this study was ignored during the same years in which Cyril Burt attained a knighthood for supposedly proving that IQ scores are almost entirely genetically determined.

A fourth argument for affirmative action is that in training a professional, we are also training a community leader. It is reasonable to train leaders of communities most in need of leaders -- notably minority communities.

Other arguments for affirmative action state that we must make up for past discrimination against certain minorities. Although I have made some of these arguments in the past, I am deemphasizing them now, not because they are necessarily wrong, but because they may be irrelevant. In the past five years, more people realize that a college or graduate degree is not necessarily a ticket to a good life or even to a job. In purely monetary terms, it is no longer even clear that a college education is a worthwhile investment of one's time and money. Currently, there is one major exception to this statement -- medical school. However, the U. S. is rapidly expanding the number of physicians it trains, so that this situation may change, and some physicians may even have trouble finding employment. A few years ago, it would have been hard to imagine an excess of lawyers or college professors, but that is what we have now. If this occurs to physicians -- and it seems likely -- then there will be no type of school in the country where admission will provide, to a member of either a majority or a minority culture, an automatic ticket to affluence and a good life. This will focus the questions of admissions policies where they belong -- on the benefits accrued to society as a whole rather than to the individual applicants admitted.

To review, I have mentioned four principal arguments for affirmative action -- the need for professionals to serve the minority community, the need for community leaders in minority communities, doubts about what aptitude tests really measure, and the idea that colleges might reasonably give lower priority to applicants who are already contributing to society in other ways. Having concluded that affirmative action should continue, and probably will, we now turn to the second major question: How should affirmative programs be managed?

After considerable debate among psychometric technicians in the last five years or so, there is now wide agreement that no mechanical psychometric formula can tell us exactly how many minority applicants to admit under an affirmative action program. Thus, these important decisions must be made by human judgment.

Who will, or should, make these decisions? Important as this question is, it is outside my competence to answer this question. Therefore, I will consider how the decisions might be made most intelligently by those who make them.

Affirmative action introduces two sets of decision makers into the admissions process, where formerly there was only one. Under affirmative action one set of decision-makers -- for instance, the courts -- essentially decides, at least in outline, how many minority candidates will be admitted. These people I call the policy makers. A second set of decision-makers reads application folders and decides on individual applicants. These people I call the policy implementers. Obviously, it is impractical to have these two sets of decisions made by the same people. So the central mechanical question of affirmative action is how the policy makers and policy implementers can communicate with each other.

Up to now, the problem has been minor because the policy makers and policy implementers have had a close working relationship. The policy makers have typically been the higher officials of a university, while the implementers have been, of course, the university's admissions officials. This practice, however, essentially excludes community participation in any part of the decision-making process.

To broaden the policy making process, we need a mechanism whereby admissions officials can release enough information about the admissions process to the public so that a broader group can have some sort of intelligent input into policy decisions, without violating the individual applicant's right to privacy. When we talk about releasing summary information without releasing information about individual applicants, summary statistics, like mean success rates or mean test scores for the different cultural groups come to mind. However, means are not really useful for this purpose since we really want to release information about the extreme applicants -- what would happen if we admitted five fewer, or five more minority applicants, rather than the present number?

In a paper published in the Journal of Educational Measurement for Spring 1976, I described, in some detail, how information can be released on the number of minority applicants admitted, without violating the privacy of individual applicants. Essentially, the technique involves pairing each minority applicant with the majority applicant who will be rejected if the minority applicant is admitted and publishing an overall acceptability score for each person. Then, anyone can see exactly what is being gained and what is being lost by admitting each minority applicant.

In summary, I have made two main points. The first is that affirmative action programs should continue. The second is that there are statistical techniques available which will enable institutions to release detailed information about the nature of the admissions process and the quality of majority and minority applicants, without violating the need for confidentiality of individual applicants. These techniques would allow more groups to have some say in determining affirmative action policies.

Let me now turn to the third, and last, portion of this paper. This concerns the possibility that we may be able to provide the applicants themselves with enough information about their chances of succeeding in any given school or college, so that they can make their own decisions, thereby allowing many schools to do away with selective admissions procedures and aptitude tests. This is a very exciting possibility, and the pioneering work is centered at the very institution -- the Educational Testing Service -- which has frequently been accused of embodying establishment conservatism. The work, done primarily by Lila Norris, Warren Chapman, and Donald Cockran is called System of Interactive Guidance and Information (SIGI).

The central concept of SIGI is that the student, rather than the institution, makes the crucial decisions about what schools to attend and what courses to take. The sophisticated psychometric information which, in the past, has been made available only to institutions would, in the future, be made available only to the individual applicants.

For years, colleges have made widespread use of a statistical technique called multiple regression, which organizes scores on several predictors -- for instance, aptitude test scores and previous grade-point averages -- into a composite prediction of the student's chance of success in the college. The SIGI system uses this same multiple regression method. However, since the results of the method are being distributed to students and not to the institutions, the information fed into the statistical analysis can include information which students would be unwilling to provide to the institution. For instance, students may answer questions about their interests in course materials or questions about their abilities and willingness to complete homework assignments on time. Under the SIGI system, this information goes directly into the computer, and nobody except the student sees it. That information is then used to provide students with the best possible estimate of their chances of succeeding in the course or the college in question.

In general, any information which in the past has been available to colleges can be made available to students, except for letters of recommendation, which are of questionable value anyway. However, as the previous examples show, there is a good deal of information which students never make available to the college. Thus, turning students, rather than institutions, into the decision-makers means that the decision-makers will have much more information than before. In short, the SIGI system essentially provides the new decision-makers -- the students -- with the sophisticated multiple-regression techniques for data-summarizing and prediction that in the past have been available only to the selecting institutions.

The studies done so far show that, in general, the new system is at least as accurate in predicting grades in individual courses as previous indicators. Results like these suggest that we can take most questions of college admissions out of the hands of the colleges and put them into the hands of the students. This would make open-door admissions policies practical in many more institutions than they are today.

Another social factor encouraging increased use of open-door admissions policies is the factor I mentioned earlier — that our nation seems to be moving toward a state in which admission to college, graduate school, or professional school no longer provides a ticket to affluence. As this becomes clearer to students and their parents, admission pressures on our select schools will decrease and greater use will be made of other institutions.

References

Darlington, Richard B. A defense of "rational" personnel selection, and two methods. *Journal of Educational Measurement*, 1976, 13, 43-52.

Gordon, Hugh. Mental and scholastic tests among retarded children. London Board of Education, Educational Pamphlet #44, published under the authority of His Majesty's Stationery Office, 1923. Excerpted in *Historical readings in developmental psychology*. Wayne Dennis, Ed. New York: Appleton-Century-Crofts, 1972.

Linn, Robert L. Fair test use in selection. *Review of Educational Research*, 1973, *43*, 139-161.

Norris, Lila. The SIGI prediction system: predicting college grades with and without tests. RB-76-26. Princeton, Educational Testing Service, August 1976.

Norris, Lila and Chapman, Warren. A test-free approach to prediction for guidance. RB-76-32. Princeton, Educational Testing Service, November 1976.

CRITICAL CONCERNS IN THE TESTING OF MINORITIES: TIME FOR NEW INITIATIVES

Ronald J. Samuda
Faculty of Education
Queen's University, Canada

The central issue in psychological testing today concerns a choice of alternatives. Bluntly stated, will testing continue to be the servant of institutional and structural racism or will it be a means of opening up educational opportunity for all children regardless of their ethnic and socioeconomic backgrounds?

In the continuing debate on the nature and use of psychological tests there is, in my opinion, an ever-present danger of missing the wood for the trees: it is indeed tempting to indulge the psychometrician's fascination for technical detail and to somehow lose sight of the larger and more fundamental perspective. So let me, at the risk of seeming to state the obvious, outline what I consider this perspective to be. The purpose or ultimate goal in any aspect of education should be the provision and improvement of learning facilities for all students. Assessment, if it is to be educationally viable or supportive, must adhere to the principle that it must serve the good of individuals by augmenting the effectiveness of the teaching-learning process. By the same token, where assessment can be demonstrated as detrimental to the teaching-learning process, it necessarily abrogates the principle of social justice and, ipso facto, contributes to the violation of the civil rights of individual citizens.

Overwhelming evidence exists that psychological tests have a deleterious effect upon certain minorities, both in terms of their educational effects and the social and economic consequences flowing from such a conditioning process. Psychological tests - specifically IQ and certain tests of aptitude - are often used to label, stultify, and retard the aspirations of blacks, Mexican-Americans, Puerto-Ricans, and native Americans. Testing, in effect, bolsters a predominant philosophy of racism endemic in the puritan ethic and implanted in the curricular structure of the schools, teacher training institutions, and attitudes espoused by text book writers, as well as test developers.

Definition

Before proceeding further, I wish to clarify what I mean by "racism" and, in particular, the two manifestations of it alluded to in my opening statement. The essence of racism, to my mind, is an assumption of inherently superior and inferior racial stocks, unequally endowed with various "human" attributes - intelligence being one of them. In action, racism attempts to scientifically legitimize this fallacious assumption and seeks to promote and maintain the domination of the racial group deemed "superior" over others deemed "inferior."

Institutional racism, of course, does not necessarily imply any personal prejudice on the part of the individual carrying out a racist policy. He acts essentially as an agent or functionary, executing what he perceives as his organizational role within the framework of a racist institution or society. (The effects of his action, however, are no less damaging to those who experience it.)

"Structural racism" is perhaps even less obvious. Yet its effects are most pervasive. "Structural racism" refers to inequalities rooted in the system-wide operation of a society which exclude substantial numbers of particular ethnic groups from significant participation in society's major social institutions. What is at issue here is not equal opportunity for those with equal qualifications, but, more importantly, the question of access of members of particular ethnic groups to the very qualifications (skills, resources) required by the majority ethnic group or groups for full participation in the life of society. Structural racism is characteristically non-intentional and impersonal in action. It has, nonetheless, an even greater discriminatory effect for the minority member who aspires to effective participation in society than does individual or institutional racism. (Hughes and Kallen, 1974, pp. 106-108)

Consequences of Testing for Minority Group Children

The implications and consequences of testing for minority individuals are very real, drastic, and far reaching in their effects at all ages and stages in the lives of minority individuals. Such profound consequences are apparent in the process of grouping and placement of children into slow learner tracks, in labelling of children as mentally retarded, in difficulties and hurdles placed in the way of those aspiring to attend institutions of higher learning.

IQ and aptitude scores are used to place students according to their perceived ability levels, resulting in the grouping of disproportionately large numbers of minority students in slow tracks or in special EMR or MR classes. This fact is well documented. Coleman's study (1966), for example, showed that at the elementary and secondary level, the schools attended by the average black child contain a significantly greater proportion of children in the low tracks. Dunn (1968) noted that at the national level, minorities comprise more than fifty percent of the mentally retarded. The figures issued by the Inter-group Relations of the California State Department of Education revealed that in 1970 blacks, who represented only 9.1 percent of the total school population of the state, accounted for 27.5 percent of the educable mentally retarded and only 2.5 percent of the mentally gifted. When we examine the effects of grouping, the significance of this sort of data becomes clearer.

Grouping

Despite the widespread preference of teachers for ability grouping as a strategy, recent surveys (Findley and Bryan, 1971) have shown that the slight effect in favour of high ability groups is offset by the negative effects on lower achieving groups. Whatever gain exists in grouping students according to ability accrues to the brighter groups only. But for blacks and other minorities, the effects of grouping are even more pernicious. As was unequivocally demonstrated in the case of Hobson v. Hansen (1967), ability grouping inevitably leads to ethnic and socioeconomic separation and thus, to racial imbalance in the classroom and the perceived "inherent intellectual inferiority" of blacks in particular. Testing, therefore, facilitates the stratification of students in terms of estimated potential, which leads to labelling, lowered self-concept, as reinforced myths and stereotypes, and lowered expectations from teachers. This, in turn, leads to poor learning conditions and the inevitable inferior performance of those students on tests of achievement. Thus, the vicious circle is formed, and the celebrated predictive validity of the IQ and other tests of intellectual aptitude is "demonstrated."

Mislabelling

Two recent pieces of evidence have dramatically illustrated the way in which test scores conspire to cause the consistent and persistent mislabelling of minority children.

Mercer's eight-year study of mental retardation, conducted in Riverside, California, shows that of the children she studied (predominantly Mexican-American and black) 75 percent of those classified as mentally retarded had been mislabelled and incorrectly placed; hence, they had suffered stigmatization and adverse consequences of a minimally stimulating learning environment (Mercer, 1971). Mercer's contention is that so-called intelligence tests, as presently used, are to a large extent "Anglocentric," that is, they mirror only the standards, values, and experiences of the white Anglo-Saxon middle-class. Consequently, whenever those tests are used with populations that do not correspond to the norm, four times as many Mexican-Americans and twice as many blacks are enrolled in retarded classes than would be expected from their proportion in the state of California. However, when a "two-dimensional" definition of mental retardation is used, a definition which not only takes into consideration intellectual performance on tests but adaptive behavior as well, and when IQ scores are interpreted with the knowledge that sociocultural factors contaminate them, then racial imbalance in classes for the mentally retarded disappears. The essence of Mercer's solution is a "pluralistic evaluation" of test results, evaluating the intelligence of each person only in relation to others from a similar sociocultural background who have had approximately the same degree of exposure to the sort of tasks required on the IQ test.

The lawsuit Larry P. et al. v. Wilson Riles et al. (1972-1977) is another case in point. This long-standing litigation has revealed considerable evidence of an alarmingly disproportionate number of black students in EMR classes in California. Significantly, these students were placed in EMR classes on the basis of IQ test scores alone. It should be noted, however, that when the plaintiffs - six black students - were retested by certified black psychologists, who used techniques which took into account the cultural and experiential backgrounds of those students, all achieved above the cut-off score of 75 which represents the borderline between mental retardates and normals. At the outset of the case in 1971, more than 27 percent of children in programs for the mentally retarded were black - although blacks comprised only 9.1 percent of the school children in California. Even today the situation has not been fundamentally altered.

Access to College

Sorting and selecting practices in the elementary and secondary schools predetermine that only a very small proportion of minority students will be eligible for admission to college. Competition further reduces this number, for often the minority student has been denied the requisite educational stimulation in school and consequently, is out-distanced by his white counterpart on the SAT. In a succinct survey of the principal barriers to higher education encountered by minorities, Crossland shows that performance on the SAT is by far the most critical (Crossland, 1971). Although many colleges have begun to use additional evidence to supplement "board scores," the fact remains that the sorting and selective function of tests still constitutes a major barrier to the higher education of minorities.

The data which follows illustrates the enormity of the problem. In 1970, minorities (namely, black Americans, Mexican-Americans, Puerto Ricans, and American Indians) were estimated to number 30.5 million, or 15 percent of the total population of the United States. Of that number, only 544,000 or 6.8 percent of the total college population were admitted to institutions of higher learning. In the same year, whites and all others, totalling 174.3 million, or 85 percent of the total population, sent 7.5 million or 93.2 percent of the total eligible college students to colleges (Crossland, 1971).

The Minority Gap-Deficit Interpretations

Faced with evidence of its harmful effects, apologists for testing insist that the fault lies within minority groups themselves - perhaps not in their stars, but certainly in their culture or their genes. After all, the argument goes, as studies of blacks' test scores show (Shuey, 1958, 1966; Dreger and Miller, 1960, 1968), minority students as a group consistently perform well below the standard of whites. It is unfortunate, they say, but the fact must be faced: minority students simply do not have the ability to cope.

Present constraints of space and time do not permit a full discussion. I will, however, attempt a brief summary of the fallacies inherent in positions of the "cultural deprivation" theorists and the "genetic deficit" adherents.

The notion of "cultural deprivation" (Bereiter *et al.*, 1966; Deutsch *et al.* 1967) and "verbal deprivation" (Bereiter and Englemann, 1966) have lost much of their respectability in the light of critical appraisal (see, for example, Baratz and Baratz, 1970; Labov, 1971; Clark and Plotkin, 1972), but many teachers and school administrators still share these views. "Cultural deprivation," apart from being a contradiction in terms, betrays an essentially ethnocentric stance which assumes that one culture is correct and that deviance from this norm constitutes social pathology. Thus, for example, a black whose lifestyle differs from that of the WASP middle-class individual is perceived as a "sick white man"

(Baratz and Baratz, 1970). Such notions at best lead to a gross misdirection of educational effort; more often, they encourage complacency within the schools and remove the burden of accountability from teachers and administrators.

The "genetic deficit" viewpoint is simple: in general, minority individuals - blacks in particular - are, in terms of intellectual ability, genetically inferior to whites (Jensen, 1969). There are several methodological and conceptual weaknesses in the research upon which this conclusion is based (Anastasi, 1976, pp. 350-353; Samuda, 1975, pp. 39-42, 50-59); only three will be cited here. First, if the genetic component in intelligence is to be properly estimated, then this component must be effectively isolated from environmental elements. This Jensen patently fails to do - and none of his supporters has succeeded in this either. If the genetic contribution cannot be estimated with any degree of accuracy, how then can one say whether one group is genetically superior to another? The second point is, Jensen's estimate of an 80 percent genetic contribution to intelligence is not only spurious but educationally irrelevant: for even if IQ has a large genetic component, it is nonetheless subject to considerable change through environmental intervention. The teacher still has a responsibility to provide the child with educational experiences of the very best quality. Thirdly, the underlying assumption that IQ tests give an accurate and reliable measure of the intellectual capacity of persons, from all cultural groups, is plainly insupportable.

The Stark Facts About IQ and Scholastic Aptitude Tests

The realities are strikingly different from what the deficit theorists would have us believe. In the first place, IQ tests and tests of scholastic aptitude are, in large measure, tests of achievement. This is not to say that they do not measure "intelligence" at all, but to the extent that they do, it is "intelligence" beamed through the prism of one group's experiences, conditioned, and deflected into a repertoire of behaviors consonant with the norms of that group. Let us take as examples two widely-used tests - the Stanford Binet and the Wechsler Intelligence Scale for Children. No child can do well on either of these tests unless he has the cultural and educational experiences assumed by the test designers, plainly evident in the language, situations, analogies, illustrations, and conceptions contained in these instruments. However, not all children have equal access to these experiences: the long-standing pattern of ethnic segregation and unequal educational provision are established facts of American history. Since the experiences tapped by standardized tests are those of the WASP middle-class culture, what these tests really measure is not the minority individual's intellectual capacity but the extent to which he has become assimilated into the mainstream culture. Therefore, to the extent that he is removed from the mainstream culture, he is programmed to fail. What IQ and scholastic aptitude tests ensure is that the existing distance from the mainstream and, along with it, exclusion from the benefits accruing to those who can participate in that culture will be maintained.

Herein lies the crux of the matter. Such tests do not recognize and take into account cultural difference (the ommission - until very recently - of minority persons from standardization samples well illustrates this insensitivity and ethnocentric bias); instead, they help to identify difference as deviance, label and penalize those who exhibit it, without providing the means by which such persons may reach the arbitrarily chosen ideal. Thus, tests function as effective instruments of structural racism.

The major weakness of intelligence testing lies in its limited educational value. It assumes that sorting and slotting are the same as educating and that an IQ score tells the teacher anything meaningful. One may, for example, infer certain learning difficulties from a given set of intelligence test scores,but the scores are of little value in identifying the nature of the problem or the steps by which it can be remedied. Indeed, the norm-referenced test, with its focus on comparison with the norm group, is irrelevant to the primary function of the school, i.e. to promote the intellectual and social growth of the individual child.

Prediction is of little value, particularly at the elementary and secondary levels. (In any case, the educational treatment that follows testing ensures the result predicted.) New effort must be directed towards developing the potential of criterion-referenced tests as diagnostic tools.

Testing can serve a useful purpose, however, if it will help to ensure the delivery of quality education to all students. To achieve this a new focus is needed - away from the traditional and wasteful task of counting sheep and goats to the more productive and challenging enterprise of developing useful human beings.

References

Baratz, S.S. and Baratz, J.C. Early childhood intervention: The social science base of institutional racism. Harvard Educational Review, 1970, 40, 29-50.

Bereiter, C., et al. An academically oriented pre-school for culturally deprived children. In F.M. Hechinger (Ed.), Preschool education today. New York: Doubleday, 1966, 105-135.

Bernstein, B. Language and social class. British Journal of Sociology, 1960, 11, 271-276.

Clark, K.B. and Plotkin, L. A review of the issues and literature of cultural deprivation theory. In K.B. Clark (Ed.), The educationally deprived. New York: Metropolitan Applied Research Center, 1972, 46-73.

Coleman, J.S. et al. Equality of educational opportunity. Washington, D.C.: U.S. Department of Health, Education and Welfare, 1966.

Crossland, F.E. Minority access to college. New York: Shocken Books, 1971.

Deutsch, M. Minority groups and class status as related to social and personality factors in scholastic achievement. In M. Deutsch et al. (Eds.), The disadvantaged child. New York: Basic Books, 1967, 89-131.

Dreger, R.M. and Miller, K.S. Comparative psychological studies of negroes and whites in the United States: 1943-1958. Psychological Bulletin, 1960, 57, 361-402.

Dreger, R.M. and Miller, K.S. Comparative psychological studies of negroes and whites in the United States: 1959-1965. Psychological Bulletin, 1968. (Monograph Supplement 70, No. 3, part 2).

Dunn, L.M. Special education for the mildly retarded. Is much of it justifiable? Exceptional Children, 1968, 35, 5-22.

Findley, W.G. and Bryan, M. Ability grouping: 1970. Status, impact, and alternatives. Athens, Ga.: Center for Educational Improvement, University of Georgia, 1971.

Hughes, D.R. and Kallen, E. *The anatomy of racism: Canadian dimensions*. Montreal: Harvest House Ltd., 1974.

Jensen, A.R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, 39, 1-123.

Labov, W. Academic ignorance and black intelligence. *The Atlantic Monthly*, 1971, 59-67.

Mercer, J.R. Institutionalized anglocentrism: Labelling mental retardates in the public schools, in P. Orleans and W. Russell (Eds.), *Race, change and urban society*. Los Angeles: Sage Publications, 1971.

Riessman, F. *The culturally deprived child*. New York: Harper and Row, 1962.

Samuda, R.J. *Psychological testing of American minorities: issues and consequences*. New York: Dodd, Mead and Company, 1975.

Shuey, A.M. The testing of negro intelligence. Lynchburg, Va.: J.P. Bell, 1958. Second edition, New York: Social Science Press, 1966.

Court Cases

Hobson vs. Hansen, 269 F. Supp. 401 (1967).

Larry P. et *al*. vs. Wilson Riles et *al*. No. C-71 2270 RFP, U. S. District Court, Northern District of California (June, 1972-1977).

TESTING MINORITY CHILDREN: WHY, HOW, AND WITH WHAT EFFECTS?

Sandra Scarr
Professor of Psychology
Yale University

To speak to issues of why, or why not, to test and how, one must examine briefly the nature of the overt and covert arguments on both sides. Unfortunately, "discussions" of the pros and cons of testing minority children are most often conducted in a Tower of Babel, where different languages and assumptions prevent productive discourse. Overtly, the opponents of the use of standard psychological and educational tests with minority children object primarily to the negative social outcomes of testing, for which they blame biases in the tests and testing procedures. Proponents of the use of standard tests defend the equal predictive validity of the measures for all groups. Some opponents of tests propose their elimination; others hope for culturally-relative measures that will eliminate average group differences. Test supporters caution about the subjective and potentially more biased nature of other assessments and the lack of predictive use for ethnically-adjusted scores. As Messick and Anderson have noted, the complaints of the testing opponents are not answered by the proponents; their respective arguments fly by like shells in the din of battle.

Covertly, two unjustified suspicions inflame the conflict about testing minority children: (1) that tests measure fixed, innate abilities, and (2) that minority children have less of this gene-given ability than majority children. I will, therefore, review briefly my own research to show that both fears are unfounded. Tests sample what has been learned, especially in relation to what is required by current school programs; and as my research shows, there is no evidence in the mean test score differences of genetic differences among ethnic groups.

If these two unfounded suspicions are allayed, I hope that a more rational look at the nature of standard tests can lead to their constructive use, and to the elimination of misuses that seem so common in current practice.

The Social and Philosophical Context of Testing

Testing on a broad scale was adopted by the schools to improve the opportunities of lower SES and minority children for selection into educationally-and occupationally-advantaged positions and to reduce the pervasive class and ethnic biases of personal judgments. The incorporation of standard tests into the schools was an example of the American ethic of equality of opportunity. But one has to be very careful about the meaning of "equal opportunity"; it is a slippery term. The use of standard tests to give all children the same chance to compete is a case of equal opportunity construed as identical treatment for all, regardless of their initial differences in preparation.

It has always been clear, of course, that minority children were not equally prepared to compete, whether they were the Irish of the 1900's, the Southern and Eastern Europeans of the 1910's and '20's, or the blacks and Spanish-speaking groups of the mid-century. We should not ignore the upward mobility that has resulted for some minority children through test scores and educational opportunities, but the overwhelming fact is that minority children have, as a group, always scored lower on tests and performed more poorly in schools before the economic assimilation of their groups into American life (Sowell, in press).

The use of standard ability and achievement tests in schools was primarily motivated by liberal views of equal opportunity as identical treatment. There are at least two other ways to construe the ethic of equality: equality of opportunity can also mean "different strokes for different folks," a view that recognizes the advantages and disadvantages that individuals bring to the competition, and equality of opportunity can be stretched to mean the assurance of equal outcomes, the selection of equal proportions of all groups (however defined) into good and bad outcomes. For the use of standard tests, the second view of equal opportunity means different investments in and preparations for the eventual competition to meet the same standards. The third view implies different standards to insure equal outcomes.

I introduce the different meanings of the ethic of equal opportunity because current arguments about the use of standard tests revolve around these underlying differences in the philosophical meanings of the term. Opponents of the use of standard tests with minority children object primarily to the negative social outcomes in the differential selection of minority and majority children into socially desirable and undesirable life tracks. They argue that the abolition of tests and the adoption of ethnic quotas will improve equal opportunity construed as equal outcomes (meaning #3). Proponents of the use of standard tests base their support on the equal predictive validity of the measures for all groups, thereby interpreting equal opportunity as identical treatment (#1). Some middle ground of meaning is occupied by those who support "different strokes for different folks" to mean that truly equal opportunity for all requires more instructional and social investment in some who begin the competition with fewer chances of success. Eventually, however, all individuals must be held to the same standard of competitive achievement (#2).

Recognizing the different meanings of equal opportunity for the various opponents and proponents of standard tests does not resolve the problem, but it puts everyone on notice that they had better be prepared to defend their meaning of equality on a broader social scale than mere psychological testing. The use of standard tests in the schools cannot be attacked or defended without a correlated justification for one's view of social equity in all spheres of educational and economic life.

My own position is unequivocally in favor of "different strokes for different folks," by which I mean that individual differences must be taken into account in our investment in and training for all spheres of educational and economic life. Some individuals need more investment of time and resources to reach some Rawlsian lowest-acceptable-level of competence. We, as a society, must guarantee that investment. In addition, differences in life chances, be they socioeconomic disadvantage, cultural difference, or genetic endowment, should be compensated to assure that every individual has the most equal possible chance to achieve; this tenet of my value system calls directly for matching instruction to individual differences, whatever the person's group membership. The measures of achievement, however, are universally applied.

This brief statement of personal principles is included to alert you to the assumptions of the paper.

Why Should We Test Minority Children?

In my view, answers to questions of why and how to test minority children are predicated on the assumption that testing should inform instruction. The proper use of tests is to provide diagnostic information to teachers that alters instruction and improves the match of instruction to the child. On a broader scale, testing can inform the recruitment of children into educationally-appropriate programs and settings that fit their current needs. Testing need not be used as an accomplice to biased selection procedures, but neither will children benefit from the denial that their school-related skills and instructional needs differ on the average by ethnic group. There are humane reasons to use tests appropriately *in the interests of the child*, whatever his ethnicity.

Covert Suspicion: Tests as Measures of Innate Intelligence

What we test as "intelligence" is based on the summation of an individual's learning experiences which prepare him to tackle new problems. All tests of intelligence presuppose past learning. Some items simply call for past learning; others require the use of past learning to solve a new problem. There is no intelligence without experience!

To use past learning as a basis for new problem solving does not imply that all individuals of the same chronological age have learned or are capable of learning the same quantity or quality of material at the same rate (genetic and biological differences exist in intelligence), but it does imply that what is tested as intelligence is not some magical "innate capacity" that is unaffected by experience. Without learning opportunities, we would all be equally stupid and ignorant. With equal learning opportunities, individuals vary enormously in what they have learned and in what skills they bring to the solution of new problems.

The Assumptions of IQ Testing

Alfred Binet was well aware that his practical device rested on several critical assumptions -- assumptions that have too often been forgotten by those who followed him. The two most important assumptions are: (1) common experiences for those being compared on the test; (2) the test as a sample of intelligent behavior.

Common Experiences

Today we recognize a profound problem in the use of intelligence tests with culturally-different populations. The information sampled on IQ tests is general cultural knowledge. But whose culture -- white, urban, middle class culture? The skills sampled are those most likely to be taught by middle-class parents.

The Test as a Sample of Intelligent Behavior

Many IQ test items seem arbitrary and ridiculous. Why should tracing a maze, arranging a sequence of pictures, knowing the capital of Greece, and defining the word "ballast" be used to measure intelligence? Tests are samples of information and skills, not complete inventories. The sampling rationale is that an individual who knows the capital of Greece is also more likely to have other kinds of rare information; a person who can repeat six digits backwards can also manipulate other information in his head, and a person who can abstract similarities between farming and manufacturing also knows how to think abstractly about other problems.

An obvious problem for minority children is the content of the test. Items are written by educated, urban, middle-class psychologists and educators to predict middle-class standards in schools and jobs. While the tests may predict well, they may not be fair samples of intelligence. At this point, we must define what we mean by "intelligence." An alternative definition is adaptability in "real-life" (not school or job) situations, such as neighborhood and family settings. Although sufficient space does not exist here to do justice to the issue of test validity, no one to date has devised a satisfactory substitute test of intelligence based on alternative criteria.

Part of the problem is the lack of agreement on what behaviors one would call intelligent in family and neighborhood settings. Is getting any job in a Black ghetto a good criterion? How about knowing how to milk a herd of cows in 30 minutes or finding your way home over ten miles of woods in West Virginia? The majority culture is not primarily interested in these feats of adaptation; we do not reward them with notable success or acclaim, but few would deny that they require "intelligence," broadly conceived.

To appreciate the difficulty in finding agreeable, alternative criteria, we have to recall the history of IQ tests; they were designed to select children who would not profit from the usual school curriculum, at a time when all children had to be served by the school system. Intelligence, as measured by IQ tests, is primarily school-learning ability, not general adaptation to life.

How Should We Test Minority Children?

Given that tests are samples of learned information and skills and that they predict well school performance for all groups, I will argue for a re-education of school personnel and the public on the meaning of test scores, the elimination of any lingering suspicion about genetic racial differences in IQ, and the abandonment of culturally-relative assessment in favor of culturally-specific assessment.

To understand why minority children, on the average, score lower on all forms of standard tests, achievement, aptitude, and IQ varieties, school personnel must be educated to see that test scores only measure past learning of information and skills that are sampled by the tests and the schools, not as more global measures of intellectual functioning for minority children. And they must come to see that children's performance can change, particularly on those achievement measures most closely tied to instructional programs. It is not enough for academic psychologists and test writers to proclaim the correct interpretation of a test score as a current assessment of what a child has learned in relation to his peers or some criterion performance.

For a minority child, this means that unless his knowledge or skills get a big environmental boost, his low scores in first grade will predict his low scores in fourth grade. Since he has the same sociocultural environment, that is hardly surprising. But suppose the schools came up with a Big Boost program for the early grades. Presumably, both his school performance and his standard test scores would change upward.

Teachers and school psychologists must be educated to understand that a test score is not written in the child's brain. More importantly, they must understand that low scores are indicators for action -- instructional action to fit the educational needs of the child. They are not excuses for labeling and discard.

The "Good Guys": Diagnostic and Criterion Referenced Tests

One does not have to extoll the virtues of diagnostic and criterion-referenced tests. Nearly everyone in education is (or should be) in favor of diagnostic tests that inform teachers of the individual child's particular strengths and weaknesses in skill development, often without comparing him to any other children. Suzy has a 60 word sight vocabulary but needs help in word-attack skills. Jimmy's phonics are okay, but his reading speed could probably be improved by instruction in phrasing and reading for meaning. Johnny understands the concepts of addition and subtraction but needs help on the concepts of multiplication and division. Surely, we would all agree that this information is useful to immediate instruction. The lack of a normative basis for the instructional decisions about what to teach next to an individual child is no hindrance.

One clear advantage of criterion-referenced tests is that they require teachers to specify the objectives of instruction and to use the test items to judge whether or not the child has achieved a specific objective. Does Suzy understand and can she use the concepts, "across" and "through"? Can Bill add two-digit numbers with carrying? If the child has not yet achieved that skill, the teacher is informed by the test to improve or increase her instruction of that skill.

One difficulty with such specificity is that not every criterion of an education can be so specified. For many reasons, test items must sample what a child knows and can do, not attempt an exhaustive inventory. Three problems that I will briefly mention are the expanding universe of knowledge and skills with development, the substitution of thresholds for quantification, and the dilemma of incomparable measurements.

In the early grades, the skills required for basic literacy can be relatively well specified, compared to those broader qualities of thinking and knowledge at later periods of development. To my knowledge, no one has tried to analyze all of the tasks involved in the interpretation of poetry or the development of a critical point of view about complex historical or social issues. Although there is still much excitement in psychology about the possibilities of specifying and organizing the learning steps in higher-level skills and knowledge, the only partially successful programming models are for young children's thinking and simple mental processes, neither of which describe or prescribe the instruction of literary or historical thought. Of course, you may object, computers can be programmed to play rather good chess, but that too is a limited domain with a finite number of possible moves, unlike most of the higher level skills I am citing. I think it is safe to say that the prescription of instructional objectives is far more feasible for second than tenth graders, unless one is still prescribing the same primary-level objectives!

Which brings me to the substitution of thresholds for quantification of the differences in skills and knowledge that children have learned. It is important to specify minimal criteria for achievements that nearly every child can and should attain. But this minimal threshold of literacy is no substitute for quantifying the remaining range of individual differences in achievements and designing appropriate instruction for all children, even those whose skills and knowledge far exceed the average of their age group. Criterion-referenced tests are simply not appropriate for this measurement problem, particularly at the older grade levels.

The final problem that I want to mention briefly is the incomparability of criterion measure from one setting to another. If we are serious about setting instructional objectives and using criterion-referenced tests to judge our effectiveness in helping children to reach those objectives, then the test items must be very closely tied to the curriculum. Because different schools teach different materials, in different orders, once a child is past the first few grades, how can tests be comparable from one school district, or building, to another. If South High in Oshkosh spends the fall semester of the eleventh grade on the poetry of e.e. cummings, a reasonable set of criterion items would test the student's comprehension, interpretation,and appreciation of that poetry. But Wilson High in Madison used the essays of Martin Luther King in their eleventh grade curriculum; the criterion items for those students will necessarily be different. I think you can see the myriad problems that arise from such diverse criteria of achievement.

The approach of many norm-referenced tests is to sample the skills and knowledge that are usually imparted in the eleventh grade. Give the student a new passage (an attempt to provide equal environmental exposure or lack of same), test his comprehension, interpretation, and appreciation of "literature." There will be some poetry samples, on which the Wilson High students will have less experience, and some prose, for which the South High youngsters will have had less exposure. Although none of the students will probably have encountered the particular test samples before, the items will assess the student's approach to literature in a gross and global way that estimates how well individual students and groups can deal with new material of the kinds they have had some exposure to in the school curriculum. The hypothetical test will be more helpful to teachers if it reports separate scores for poetry and prose, for reading comprehension, vocabulary, and the like, than if it reports a general "literature" score.

Now, it probably occurred to you that this same norm-referenced test could be used as a criterion-referenced one, by the simple expedient of reporting the item results to the teachers. The eleventh grade English teachers at South High and Wilson High would receive profiles of performance for their students, presumably with differential achievements in the poetry and prose sections according to the curricular emphasis. In addition to receiving the norms of the individual students and the class, the teacher should be able to see whether his/her instruction affected the profile of performance and to see in which areas the students need further work.

For the purposes of immediate instruction, the teacher has no need for the norms at all. In fact, given the ways that too many teachers interpret norms, their absence is probably better. But in most ways, a criterion-referenced test is merely an unstandardized one; anything a criterion-referenced test can do could be done just as well if it were standardized. Conversely, the use of standardized test items as criteria of learning transforms it into a criterion-referenced test without changing an item.

It is all very well to eliminate norm groups for tests that diagnose next instructional steps and measure instructional objectives in the immediate teaching situation. But it is much less appropriate to eliminate norms for other kinds of assessment. Not all instructional decisions are or can be immediately applied to classroom strategies with children. Some instructional decisions involve planning months or years ahead, making educated guesses about the suitability of the program for the child. Of course, these decisions are less valid and reliable than those made for the immediate situation, if the tests are equally well constructed (which of course they usually are not). But advice must be given about planning the child's curriculum for junior high or high school, planning for algebra, business math, Latin, or remedial English. Standard achievement tests are, in my knowledge, the best advice givers on these and similar matters, along with school grades and teachers' experiences with the child.

For an even more temporally remote kind of decision, selection for higher education is, I believe, most efficiently done with a combination of standard aptitude test, school grades, and recommendations. The role of test scores in this process has been challenged, but I think that a dispassionate look at the data would convince most people that global samples of school-related knowledge and general intellectual functioning in school-related skills provide the best prediction of who should be where in post-secondary education. Note that I am not making judgments about non-school related intelligence, just about who will be more likely to succeed in given instructional programs.

The "Bad Guys": Even Global Assessments Have Their Uses in the Interests of Children

Most criticism of testing has been leveled at the standard, group aptitude and individual IQ tests. The use of a single score or a few sub-test scores to make like decisions for a child has been rejected by those who note that average group differences in test scores classify more minority children as retarded. Low test scores alone have been used to place them in special classes. Clearly, this is poor practice.

Let us consider possible legitimate uses of such tests. First, minority children who score well on standard tests are the forgotten 25 to 30%, whose above-average to high scores are tickets to educational advancement, particularly in these days of affirmative action. In every school building in Philadelphia there are black children scoring in the 99th percentile on standard aptitude tests (and achievement tests as well). For some of these children,whose reading and math skills are not well developed, their high aptitude scores will alert teachers and counselors to their instructional needs and potential selection into higher education.

A simple classification of children by above-and below-average aptitude and achievement tests will quickly show that the aptitude tests give additional information. There are sizeable numbers of children with specific learning disabilities whose reading and math scores are low but whose aptitude scores reveal general abilities to do well in scholastic skills, once the specific problems are remedied. These children can be said to be "saved" by the aptitude tests from being put into inappropriate, generally slower groups.

What about those who score low on both aptitude and achievement tests? This is the source of the angriest attacks on standard tests because many schools lump these heterogeneous children into a single slow or retarded group. Undoubtedly, many of the minority children in the low scoring group should not be considered generally mentally retarded or even slow learners at non-school tasks. No matter what the causes of their slow school learning, it is not useful to label them as mentally retarded because they will not be so considered when they leave school. The probable cause of their slow school learning is cultural difference and inappropriate instruction. This does not mean, however, that they will thrive in a regular classroom with 28 other children, many of whom were reared in the majority culture.

Their low scores on standard tests should be only the first clue that their varied performances should be further evaluated to obtain a more complete picture of what they can do, not what they cannot do on standard tests. But note that the tests are the first clue that something is amiss with the match of the school program to the child with low aptitude and achievement scores. Perhaps, he cannot read well enough to take any of the tests, including math instructions. Perhaps, he is in need of individual encouragement to try harder. Perhaps, teachers have misunderstood his quietness or outbursts and turned him off entirely. Perhaps, an auditory mode of instruction is better suited to his learning needs.

In other words, the only use of low standard test scores should be to alert school personnel to the need for further individual study of the child. They should not be used alone to make any decisions about special classes or even special instruction.

What would we do without tests?

There are two major suggestions for a world without standard tests. First, we could return to the turn of the century and teacher ratings. Second, we could try everyone at everything they wanted to try and eliminate those who fail at the criterion performance. The first, I will argue, is even more prejudicial than tests, less reliable, and likely to lead to some wretched wrangling over minority quotas, in perpetuity. Subjective judgments are no replacement for objective assessments, although teachers often have valuable, additional, qualitative information about a student's work habits, interests, personality and the like. Recommendations should not supplant performance on a standard test.

The second, doing away with prediction and diagnosis and trying the criterion, may work reasonably well in immediate learning situations, but it is unlikely to be useful in selection of a few students from many applicants for higher education or desirable jobs. Those of you at CUNY know the experience of open admissions far better than I, but I have not heard of any great success stories emanating from the experiment. One likely outcome is that the exit criteria are simply adjusted downward to permit graduation for the majority who no longer can meet the old standards of excellence. When we insisted that everyone stay in high school to graduate, we found that a high school diploma now guaranteed only reasonably frequent attendance, not any minimum standard of achievement in reading, math, or cultural knowledge. Now, nationwide, the schools are rapidly returning to criteria of competency for the high school diploma, according to the New York Times. Can you guess which groups of children will suffer the highest failure rates?

Somehow, none of this is the answer for the instructional needs of minority children. To deny that we have a problem in educating minority children to the knowledge and skills of the majority culture is to hide one's head in the proverbial sand; to pretend that doing away with the tests will fix the ethnic disproportions in the various school curricula is to whistle Dixie in Harlem.

Covert Suspicion: Genetic Racial Differences in IQ

Research by genetically-oriented psychologists, including myself, has shown repeatedly that individual differences in test scores are caused in part by genetic individual differences. Does this fact have anything to do with the average test score differences among racial groups?

Three recent investigations on the possible genetic origins of racial differences in performance in school and on IQ tests reject the hypothesis of genetic differences as a major source of those differences in performance. The study of 130 transracially adopted black children (Scarr & Weinberg, 1976) showed that black and interracial children reared by socioeconomically advantaged white families score very well on standard IQ tests and on school achievement tests. Being reared in the culture of the test resulted in intellectual achievement levels for black children that were comparable to adopted white children in similar families. Therefore, it is highly unlikely that genetic differences between the races could account for the major portion of the usually-observed differences in the performance levels of the two groups.

A second study on the relation of black ancestry to intellectual skills within the black population (Scarr, Pakstis, Katz & Barker, in press) showed that having more or less African ancestry was not related to how well one scored on cognitive tests. In other words, holding constant social identity and cultural background, socially classified blacks with greater amounts of white ancestry did not score better than other blacks with more African ancestry. A strong genetic difference hypothesis cannot account for this result.

The third study was of black and white twins in Philadelphia. Briefly, the black 10 to 16 year olds scored half to one standard deviation below the whites on every cognitive measure. The social class differences between the races were not sufficiently large, as Jensen has reminded us, to account solely for the magnitude of this performance difference between the racial groups. The major hypothesis was that black children have less overall familiarity with the information and the skills being sampled by the tests and the schools. By using twins in this study, we were able to examine three implications of cultural differences compared to a genetic differences hypothesis.

We believe that the pattern of results supports a general cultural difference hypothesis far better than a genetic differences view. The major intellectual results are:

(1) Black children have lower scores on all of the cognitive tests, but they score relatively worse on the more culturally-loaded of the conceptual tests.

(2) The cognitive differences among the black children are less well explained by genetic individual differences, by age, and by social class differences than those of the white children.

(3) The similarity of the black co-twins, particularly to DZ's, suggests that being reared in different families determines more of the cognitive differences among black than white children, but that those between-family differences are not those usually measured by SES variables in the white community.

Therefore, we conclude that the results of this study support the view that black children are being reared in circumstances that give them only marginal acquaintance with the skills and the knowledge being sampled by the tests we administer. Some families in the black community encourage the development of these skills and knowledge, while others do not. In general, black children do not have the same access to these skills and knowledge as white children, which explains the lower performance of black children as a group. The hypothesis that most of the differences among the cognitive scores of black and white children are due to genetic differences between the races cannot, in our view, account for this pattern of results. Therefore, in 3 studies the hypothesis of genetic differences between the races fails to account for the IQ performance differences.

Culturally-Relative Assessment

Thus, we know that the major causes of the lower scores of black children reared in the black community are cultural differences and socioeconomic disadvantage. The mean scores of minority children are lower than they would be if they had equal access to the majority culture.

But that does not mean that the performance of minority children on other types of school tests or school grades will be less well predicted. Nor does it mean that adding points to their test scores will fix their later achievement difficulties. As many others have noted, the attacks on testing will not improve the match between instruction and the child.

What then about minority youths who on the average do not score as high on standard aptitude tests or in school grades; should they be given compensatory scores, padded to take into account their cultural differences? It may be cosmetic, but it does not answer the challenge of matching instruction to the educational needs of the minority youth. Their lower scores do predict lower post-secondary school performance and fudging them upward will not in itself help anyone to provide more appropriate teaching or program selection. It may well be true that outside of school the minority youths with low scores can do better than majority youths with similarly low scores because the lower scores of the minority youths result in large part from their culturally different rearing. But that fact will not help them in scholastic settings where the standards are majority norms. If you adjust standard test scores according to the home background of the child, you fall into the trap of defining tests scores as measures of intelligence in the most offensive way; in other words you have bought the misuse of IQ tests.

Culturally-Specific Assessments

There are good reasons, on the other hand, to develop culturally-specific, not culturally-relative assessment procedures. To assess the strengths of a minority child in terms of functioning within his own community can be useful in predicting his adaptation to that community. By culturally-specific assessments, I mean those measures that discriminate the more from the less skillful in the terms of the minority community. It is doubtful, however, that such culturally specific assessments will improve the prediction made by standardized tests to adaptation in the majority community because as the predictors vary so do the criteria. But culturally-specific assessments can remind the schools and others that the current IQ and aptitude tests are better predictors of school than life performance and are probably more closely tied to life in the majority than minority communities.

The more compelling reason I see for the development of culturally-specific assessments is to better understand the bases of achievement in minority communities, the differences in expectations between minority and majority cultures, and to use that information to help children bridge the gulf between community and school. At present, we really do not understand the variety of skills that minority children bring to school that may be equivalent to, but different from, those of the majority group, whose skills the school capitalizes upon. Something called "social competence" may capture what I mean here, but it is poorly defined at present and not well measured in any community. Whatever we call this culturally-specific set of skills, it is a fairly safe bet that the skills developed in the majority community are more fully sampled by the current school programs than any culturally-specific set of competencies from minority communities. Knowing what their differences are will help educators to capitalize more fully on what minority children can do, rather than always approaching their instruction as deficient members of the majority culture.

Toward an Informed Instruction

The norm of a year's progress for a year spent in school is often not met by minority children, and when it is, they are still behind because they came to school in kindergarten well behind their majority peers in the skills that schools expect. There is no evidence that minority children fall farther behind the majority group through the school years, once one takes into account the larger individual variation in school performance later in the school career. That is, second graders cannot be 4 years behind in reading skills, whereas eighth graders can, and often are. Once these individual variations are scaled in standard deviation units, the gap between minority and majority children is the same from kindergarten to twelfth grade. Thus, the schools do not aggravate the performance gap, but neither do they narrow it. Perhaps, one ought to ask what kind of special programs could be applied to narrow the academic gap, so well predicted by the early tests and so well recognized by kindergarten and first grade teachers.

A different kind of special instruction is required, one that is not equally applied to all children by race and color, but one that is applied to those children who need it. Equal opportunity need not be construed as equal treatment for all because if it is, it ignores the large individual differences in learning rates and does nothing to remedy the performance gaps between groups. If we want to give more minority children an equal opportunity, we will have to give them more time and instruction in the school skills that will make them literate citizens in the majority culture.

Perhaps, more minority children do need special instruction than children whose home cultural environment shares more features with the school culture. Perhaps, there ought to be a new category of special instruction. Call it what you like -- majority cultural instruction, the culturally-different hour, whatever. The problem with the instruction of minority children in the current schools lies not with the teachers, who as a group are usually dedicated, hard-working adults; nor with standard tests that evaluate children according to their *performance* levels, regardless of why they perform that way; nor with the children, whose task it is to comprehend and learn what the schools have to offer. The problem is with the mismatch of instructional norms and strategies to many minority children and our ignorance about the skills they do have.

FOOTNOTE

1. The controversial nature of this paper does not lend itself to traditional citations of the literature. I am especially indebted to the following articles, whose authors, however, do not necessarily share my views:

Bennett, V.D.C. Intelligence Testing, *Clinical Child Psychology Newsletter*, 1970, *9*(3), p. 8.

Gordon, E.W. Human Diversity, Program Evaluation and Pupil Assessment, *IRCD Bulletin*, 1977, *12*(1), pp. 1-7.

Gordon, R.A. Examining Labelling Theory: The Case of Mental Retardation, in Walter R. Gove (Ed.) *The Labelling of Deviance: Evaluating a Perspective*, New York: Wiley Press, 1976.

Messick, S. & Anderson, S. Educational Testing, Individual Development, and Social Responsibility, *Counseling Psychologist*, 1970, *2*(2), pp. 80-88.

Milgram, N.A. Danger: Chauvinism, Scapegoatism, and Euphemism, *Clinical Child Psychology Newsletter*, 1970, *9*(3), pp. 2-3.

Newland, T.E. Testing Minority Group Children, *Clinical Child Psychology Newsletter*, 1970, *9*(3), p. 5.

Samuda, R.J. *Psychological Testing of American Minorities: Issues and Consequences*, New York: Dodd, Mead and Company, 1975.

Wikoff, R.L. Danger: Attacks on Tests Unfair, *Clinical Child Psychology Newsletter*, 1970, *9*(3), P. 4.

Williams, R.L. Danger: Testing and Dehumanizing Black Children, *Clinical Child Psychology Newsletter*, 1970, *9*(1), pp. 5-6.

Williams, R. L. From Dehumanization to Black Intellectual Genocide: A Rejoinder, *Clinical Child Psychology Newsletter*, 1970, *9*(3), pp. 6-7.

BIBLIOGRAPHY

Anastasi, A. (Ed.) Assessment in a Pluralistic Society, *Proceedings of the 33rd Invitational Conference on Testing Problems*, Princeton, N. J.: Educational Testing Service, 1973.

Boehm, A.E. Criteria-referenced assessment for the teacher, *Teachers College Record*, 1973, *75*, 117-126.

Brim, O.G., Jr. American attitudes toward intelligence tests, *American Psychologist*, 1965, *20*, 125-130.

Caldwell, M.B. & Knight, D. The effect of Negro and white examiners on Negro intelligence test performance, *Journal of Negro Education*, 1970, *39*, 177-179.

Cleary, A.T. Test bias: Prediction of grades of Negro and white students in integrated colleges, *Journal of Educational Measurement*, 1968, *5*, 115-124.

Costello, J. Effects of pretesting and examiner characteristics on test performance of young disadvantaged children, *Proceedings of the 78th Annual Convention of the American Psychological Association*, 1970, *5*, 309-310.

Crown, P. J. The effects of race of examiner and standard vs dialect administration of the Wechsler Preschool and Primary Scale of Intelligence on the performance of Negro and white children, Doctoral Dissertation, Florida State University, Ann Arbor, Michigan: University Microfilms, 1970, No. 71-18, 356.

Bibliography, continued

Ebel, R.L. The social consequences of educational testing, School and Society, 1964, 92, 331-334.

Goslin, D.A. Standardized ability tests and testing, Science, 1968, 159, 851-855.

Holtzman, W.H. The changing world of mental measurement and its social significance, American Psychologist, 1971, 26, 546-553.

Kohlberg, L. & Zigler, E. The impact of cognitive maturity on the development of sex role attitudes, Genetic Psychology Monogram, 1967, 75, 89-165.

Sattler, J.M. Assessment of children's intelligence Philadelphia: W.B. Saunders Company, 1974.

Scarr, S. Genetic effects on human behavior: Recent family studies. A Master Lecture, delivered at the annual meeting of the American Psychological Association, San Francisco, August 26, 1977.

Scarr, S., Pakstis, A.J., Katz, S.H., & Barker, W.B. The absence of a relationship between degree of white ancestry and intellectual skills within a black population, Human Genetics, 1977, 857, 1-18.

Scarr, S. & Weinberg, R.A. IQ test performance of black children adopted by white families. American Psychologist, 1976, 31, 726-739.

Sowell, T. (Ed.) American Ethnic Groups Washington, D.C.: The Urban Institute, 1977.

Bibliography, continued

Thorndike, R.L. Concepts of culture-fairness, Journal of Educational Measurement, 1971, 8, 63-70.

Valentine, C.A. Deficit, difference, and bicultural models of Afro-American behavior, Harvard Educational Review, 1971, 41, 137-157.

Williams, E.B. Testing of the disadvantaged: New opportunities. In F.H. Wright (Chm.), Uses and abuses of psychology: A program For constructive action. Symposium presented at the American Psychological Association, Washington, D.C., September, 1971.

Zigler, E. & Butterfield, E.C. Motivational aspects of changes in IQ test performance of culturally deprived nursery school children, Child Development, 1968, 39(1), 1-14.

PART III - TESTING AND TEACHER UNIONS

WHY STANDARDIZED TESTS FAIL THE ACCOUNTABILITY TEST

Frances Quinto
Professional Associate
National Education Association

The controversy about the role of standardized testing is not a struggle between the forces of good and evil. Honest and sincere educators have real philosophical differences about testing and about the evaluation of school-based learning. Unlike many who have attacked the motivations of the NEA in opposing standardized testing, we prefer to avoid such attacks and to present the factual basis for our opposition to tests. Our position is not based on political utility, but upon long and careful consideration of the issues.

First, norm-referenced standardized tests tell us nothing about what children ought to know or what they ought to be able to do. Norm-referenced tests are specifically designed to spread children out over the range of a "normal" curve so that about half of the children in the norm group must score "below average." Grade level is simply the middle score -- half of the children in the norm group must be at grade level or below, and half must be at grade level or above. Grade level equivalency scores reflect how a national sample of students at different grade levels actually performed on the test, not how children at a particular grade level ought to perform. Specifying how children at a particular grade level ought to perform would require a statement of specific educational values and objectives in terms of learning theory and developmental psychology. A consensus among educators on issues of values, objectives, and principles of educational psychology for children at different grade levels in school is still lacking. Thus, norm-referenced tests substitute statistical comparisons in place of basic educational decisions about what children should be able to achieve. Norm-referenced test questions are statistically selected on the basis of their abiliity to spread students out over a wide comparative range not on the basis of fundamental criteria that establish what students should know. The norm-referenced testing approach to educational accountability abdicates to statisticians and psychometricians the parents' and teachers' responsibilities for establishing criteria for achievement.

Norm-referenced standardized achievement tests assess a very narrow range of student achievement. As Ralph Tyler observed, standardized tests get "small answers to small questions."[1] Tests emphasize rote recall, simple facts, and routine arithmetic. These are important educational outcomes, but they are far from the whole picture; the tests measure only a small part of the educational process. Tests frequently contain trivial items that help distribute students over a wide range but reveal little about rudimental student capabilities. The format of tests emphasizes getting the "right" answer with little attention to problem-solving processes or creativity. Higher-order thought processes are severely neglected. Indeed, in terms of the broad range of school activities and goals, standardized tests attempt to assess a minor part of what a teacher does with students. Given these limitations, in no way can scores on a single test be considered an indicator of the effects of schooling. The effects of schooling include the abilities to synthesize, integrate, analyze, make deductions, estimate, relate to others, make decisions, and develop self-esteem. Not only does the call for accountability through test scores ignore these important schooling outcomes, but those outcomes that are measured are assessed incompletely. Tests give only a partial picture of student achievement in basic skills because items that all or most students have learned are deleted from the tests in order to maximize statistical variation.

---

[1] Quoted by Bernard McKenna, "What's Wrong with Standardized Testing?" Today's Education, p. 2.

Norm-referenced standardized tests administered to groups are not useful for diagnosis of individual student needs. First, the range of cognitive areas assessed in any one test is so small that only a very few of a student's achievements are measured. Such tests provide very little information about individual students. Secondly, the information for any given student is highly unreliable for a test administered only once.

There are many sources of error. The health of the student on the day the test is given can affect the score. Whether or not the pupil had breakfast can make a difference. Noise in the classroom, a sudden fire drill, whether or not the teacher or a stranger gives the test, a broken pencil, and any number of similar disturbances can change a test score. The mental state of the student -- depression, boredom, elation, a conflict at home, a fight with another student, anxiety about the test, a low self-concept, motivation to perform -- all of these factors affect how well the student performs. Simple mechanical errors, such as marking the wrong box on the test sheet by accident, or accidentally skipping a question, or missing a word while reading are common problems for all of us. Students who have trouble reading will perform poorly on reading tests, but they are also likely to perform poorly on social studies, science and arithmetic tests because all of these tests require reading. Thus the test may considerably underestimate the real knowledge of students.

Some students perform better on tests because they have been taught how to take written tests. Some students are simply better test-takers than others because of their backgrounds or personalities or how seriously they treat the idea of the test. Some schools make students sit all day long taking test after test, sometimes for an entire week. Other schools give the tests for only a half day or two hours a day to minimize fatigue and boredom. Some students like to take tests; some do not. Some teachers help students with difficult words or even read the tests along with the students; others do not. Some schools devote their curricula, or at least some school time, to teaching students what is in the tests. Other schools, notably alternative schools -- open classrooms, free schools, street academies -- place little emphasis on test-taking and paper-and-pencil skills, thus giving students less experience in the rigor and skill of taking tests.

All of these sources of error -- and we have scarcely scratched the surface of possible sources of error -- can seriously affect an individual student's score. Moreover, these are all sources of error that have virtually nothing to do with how "good" the test is, how carefully it was prepared, and how valid its content is for a given student or group. These errors are intrinsic to the nature of standardized testing. They are largely uncontrollable errors, always present to some extent. They are the reason that statisticians can never develop a test that is 100 per cent reliable.

These errors are more or less serious depending on how a test is used. When looking at test scores for large groups it is expected that some students will perform above their true levels and others below because of errors. For most groups, statisticians believe that these errors cancel each other. The overly high scores of some students compensate for the overly low scores of other students so that the group result is relatively accurate. The larger the group tested the more likely this is true.

However, for a specific individual there are no other scores to make up for the error in that individual's score. The only hope is that the questions the student answered wrong because of error will be compensated for by the questions the student got right either accidentally or by guessing. This type of error compensation is much less reliable in correcting for error than the situation described for large groups. The least reliable result is one individual's answer on a single question. Nothing can compensate for error in this case. Thus, the tests are virtually worthless for diagnostic purposes. And as Edmund J. Farrell, of the National Council of Teachers of English, said at the 1976

Invitational Conference on Testing in Indianapolis: "Tests which are not diagnostic are educational dead ends."[2]

The most fundamental objection of all is that norm-referenced standardized tests are not valid measurements of the actual learning experiences of students. For example, in one evaluation research project Edna Shapiro found no differences in achievement test scores between (a) children in an enriched Follow Through (FT) program modeled along the lines of open education, and (b) children in comparison schools not involved in FT or other enrichment programs. When the children's responses in the test situation were compared, no differences of any consequence were found. When observations were made of the children in their classrooms, however, there were striking differences between the FT and comparison classes. The children also behaved differently in these two types of environments. Standardized achievement tests failed to detect these differences. Shapiro suggests that "there were factors operating against the demonstration of differences,"[3] which call into question traditional ways of gauging the impact and effectiveness of different kinds of school experience. The testing methodology, in fact, narrowed the nature of the questions that were being asked and predetermined nonsignificant results. She writes:

---

[2] Edmund J. Farrell, "The Inadequacies of Standardized Tests." Speech delivered at the Invitational Conference on Standardized Tests for Large School Districts, American Psychological Corp., Indianapolis, Indiana, May, 1975.

[3] Edna Shapiro, "Educational Evaluation: Rethinking the Criteria of Competence." School Review, November 1973, p. 532.

> The individual's responses in the test situation have conventionally been considered the primary means to truth about psychological functioning. Test behavior, whether considered as a sign or sample of underlying function, is treated as a pure measure. Yet the test situation is a unique interpersonal context in which what is permitted and encouraged, acceptable and unacceptable, is carefully defined, explicitly and implicitly. Responses to tests are therefore made under very special circumstances. The variables that influence the outcome are different from those which operate in the classroom, but the notion that the standard test or interview provides equal treatment for all subjects is certainly open to question.[4]

Standardized tests can bias evaluation results by imposing a standardized and controlled stimulus on an environment where learning depends on spontaneity, creativity, and freedom of expression. Shapiro found that the results of the standardized test measured response to stimulus (the test), which was essentially alien to the experience of the children. Such tests fail to delineate the learning outcomes of children who make different uses of particular classroom situations.

The problem is not simply one of finding a new or better standardized test, but of understanding the context of observed behaviors, the meaning of specific achievement outcomes of the child in a more holistic setting than is possible with any standardized test. Thus the validity of the tests is highly suspect -- they do not accurately measure what a child can achieve in natural classroom settings.

Standardized tests have also consistently been demonstrated to contain language and content that discriminate against minority group children. The tests assume a homogeneous culture. Clearly, the different cultural experiences of children affect their performances on tests. Herb Mack and Ann Cook,[5] teacher trainers in New York City, interviewed minority children about their responses on tests. They found that many children had thoughtful and reasonable responses on reading tests that were considered wrong answers on the test. The issue of cultural bias in standardized tests has been widely discussed. Members of minority groups have been particularly sensitive to these issues. The NAACP has joined the NEA in calling for a moratorium on standardized testing.[6] After careful study by a task force, the NAACP concluded that standardized tests were highly discriminatory in content and use.

---

[4] *Ibid.*, p. 534.

Edmund J. Farrell documented this point before some fifty directors of research and evaluation for large city districts in 1976. He said in part:

> In light of the cultural mosaic that is America, it is nonsense to talk about national norms against which the achievements of an individual or a group can be accurately compared. There exist only test publishers' norms, and the sample populations selected to norm tests are not always representative in time or place or race

---

[5] Herb Mack and Ann Cook direct the Community Resources Institute, Brooklyn, New York. They reported the results of their interviews at the Third Annual Meeting of the North Dakota Study Group on Evaluation, Minneapolis, MN February, 1975.

[6] The NEA moratorium resolution was passed in 1972. Since that time the NEA has moved beyond the call for a moratorium and now advocates the elimination of group-administered standardized tests. The NAACP issued a moratorium statement in May, 1975.

> of the communities of students who take the tests. As the authors of Common Sense and Testing in English note, "Many English tests have content totally alien to large groups of young people, such as reading passages about life in a Maine fishing town or about a world of butlers and housemaids on a test used in a big southwestern city with a large Latino and black student body."
>
> Items that are culturally biased against a population different from the norming population lead to the legitimate protests of those discriminated against.[7]

Standardized tests operate in several ways to submerge meaningful individual differences in learning styles, interests, and abilities, while reflecting relatively meaningless differences in children such as differential ability to take tests. Children vary in their test-taking experiences and abilities. Some students become quite proficient at test-taking. Students can also be taught how to take the tests or "taught to the test," so that their scores reflect their specific test-taking abilities, and not some 'real' performance level. Some of this results from poor testing procedures, but it is also fundamentally an inherent part of any standardized testing situation.

Standardized tests are poor instruments of accountability. At the individual student and classroom level the tests are highly unreliable and they are invalid measures of teacher performance. Noted educator and evaluator Ralph Tyler has pointed out that:

---

[7] Op. Cit.

> A standardized test is designed to be used in schools throughout the nation, despite the different learning sequences they have and with children coming from a variety of backgrounds and at various stages of learning in the field covered by the test. For this reason, it cannot include enough questions appropriate to each child's stage of development to measure reliably what he has learned during a single year.[8]

The problem is that when one seeks to find out whether a student who made a low score has learned certain things during the year, the tests do not include enough questions about work the child has actually done to furnish a dependable and representative view of a child's achievements. How, then, can a test reflect a teacher's achievements?

Another reason why standardized tests are poor accountability instruments is that any given teacher is only partially responsible for what a student learns. A fifth grade teacher may take a child who enters the class with no reading skills to a point of basic reading, but the child will still score well below national norms on a test. In no way does such a test reflect a teacher's efforts.

Furthermore, schools and teachers are not responsible for social conditions that militate against learning. School achievement is related to parental education and income, both of which in this country are unfortunately related to race. Robert Thorndike,in a study conducted under the auspices of the International Association for the Evaluation of Educational Achievement,found reading comprehension to be directly related to students' socioeconomic backgrounds. Thorndike writes in his summary of findings:

---

[8] Ralph Tyler, "Testing for Accountability," Nations Schools, December, 1970.

> A second main finding is that in the developed countries an appreciable prediction of the reading achievement of individual students -- and an even more substantial prediction of the average reading achievement of children in a school -- is provided by information about their home and family backgrounds. A dominant determiner of the outcome from a school in terms of reading performance is the input in terms of the kinds of students that go to the school. When the population of a school comes from homes in which the parents are themselves well educated, economically advantaged, and able to provide an environment in which reading materials and communications media are available, the school shows a generally superior level of reading achievement.[9]

The implication is that if the society wishes to increase achievement dramatically, it needs to provide students with well educated, reasonably affluent parents, an implication that Christopher Jencks would appreciate. In short, the results of tests are a better indicator of student background than teacher performance.

Seven major objections to group-administered standardized tests have been outlined here. (1) Norm-referenced tests tell us nothing about what children ought to know or what they ought to be able to do. Instead the tests are designed to statistically spread students out over a wide comparative range substituting statistical comparisons for basic knowledge about achievement on specified educational objectives. (2) These tests assess only a very narrow range of student achievement; higher-order thought processes and related outcomes of schooling are largely ignored. (3) The tests are virtually worthless for diagnostic purposes. (4) The tests are not valid instruments for measuring the actual learning experiences of children; observed classroom performance in natural settings is often markedly different from test performance. (5) Standardized tests inaccurately reflect the abilities of many minority children (6) Standardized tests operate to submerge meaningful individual differences in children. (7) Standardized tests are poor instruments of accountability because they reveal very little of what a teacher actually does and because the test results mainly reflect student background.

---

[9] Robert Thorndike, Reading Comprehension Education in Fifteen Countries. New York: John Wiley, 1973.

Other objections [illegible] testing which warrant further discussion are the following: (1) Tests are a negative, and often frightening, experience for students. (2) Tests frequently distort the curriculum. (3) National norm-referenced tests ignore local curriculum variations and differences in educational goals. (4) Tests work against creative students and innovative programs. (5) Tests assume a common, standard core of knowledge that can be applied to all children from all parts of the country. (6) Tests represent an enormous expense for which there is little return.

The NEA is also working to develop and implement evaluation alternatives to standardized testing. The NEA definitely supports meaningful and valid systems of educational accountability or evaluation. The NEA has published several monographs discussing evaluation alternatives, including a special publication entitled Alternatives to Standardized

Testing. The issue is not accountability itself, i.e., whether or not educators ought to be accountable. The issue is how to conduct evaluations that are meaningful and useful for improving teaching and learning. We are convinced that standardized tests do not meet these criteria.

TESTING: TEACHER UNIONS AND PROFESSIONAL RESPONSIBILITY

Irwin H. Polishook
Vice President
American Federation of Teachers

During Edward Koch's tenure as congressman and, more recently, in his campaign for mayor of New York City, he frequently met voters with the question, "How am I doing?" The technique reminds me of the patient who asks his physician, "How do I feel?" The questions are perhaps not better or worse than some of those appearing on tests given to students, but they do illustrate some of the purposes of tests in the schools.

Tests are most often used to determine the learning abilities of a student and the effectiveness of instruction received. The most constructive of these tests, and probably the least controversial, are given during the course of instruction, and they are least problematic when their grades are not recorded or do not "count." Their function is to determine not only how much the student has learned but, more particularly, the quality of what has been learned and what has not been grasped. These tests are inseparable from evaluation: telling the student what is being achieved, whether the individual's approach to a course is adequate, classroom activity is productive, study habits are efficient and how well progress is being made in relation to one's classmates.

Ideally, tests inform the student. This may be done directly or through interpretation by the teacher, revealing what deficiencies are manifest, how they may be overcome, and how to improve any future performance. For the student, the learning process and the testing procedure are functionally interrelated.

Such tests are also important to the teacher. They may indicate shortcomings in performance, hitherto undetected weaknesses in the students as a group, and selected difficulties of individual students. Tests point to ways of tailoring plans for future instruction to meet more precisely the special needs of the class -- of this class as distinguished from the previous class and the next one.

As long as tests are used for these clearly pedagogic purposes, there is little controversy over their utility. As feedback, they are part of the teaching-learning process and, in fact, occur informally, constantly, and without having negative connotations in a healthy student-teacher relationship.

Beyond these constructive purposes, the purposes of testing that are being debated in education today are those that determine a student's placement into a selected level of instruction or entry into an occupation, eligibility for promotion, graduation or credentialing, and ranking among other students in identical categories for competitive purposes that may involve further schooling, employment, and awards or scholarships. Tests certainly deserve the attention they are receiving here and elsewhere, for they have lasting effects on the lives of students and on our whole society.

Because of their critical importance, tests used to determine admission to college are especially important. Moreover, to the degree that college admission tests are problematical, they may serve as a paradigm for the purpose of understanding the relationships among tests, teacher unions, and professional responsibilities.

College admission examinations are designed not only to measure performance in secondary school but also to predict performance in postsecondary school. For predictive purposes, such tests need to factor out the distortions generated by local idiosyncrasies, geographic differentials, sociological distinctions and the other myriad influences on a student's test scores that corrupt their predictive accuracy. Failure to eliminate such extraneous factors penalizes the student by diminishing educational opportunity and thereby minimizing, if not eradicating, employment opportunity, social mobility and all the advantages that go with them. We are dealing, therefore, not only with a technical or theoretical measurement problem concerning examination construction and college admissions officers but also with a profound social problem affecting access to postsecondary education, the opportunity for careers, and the lives of millions of students.

In recent years, consciousness of the problem has been raised to perhaps exaggerated proportions by the decline in scores on the Scholastic Aptitude Test (SAT) and other examinations used widely to determine college admission. SAT scores, for example, have dropped steadily over the past fourteen years. If that were not sufficient cause for alarm, the scores seem also to reflect a cultural bias against the poor, ethnic minorities, and women who register lower-than-average records on one or both sections of the test.

A study panel appointed by the College Entrance Examination Board, which sponsors SAT, took two years to analyze the meaning of the test-score decline. The panel, headed by Willard Wirtz, found no single explanation for the phenomenon. It did put forward some contributing causes. Instances cited in the schools were the proliferation of elective as opposed to basic courses; outside the schools, the influence of television. The Wirtz study attributed the decline in test scores to the extension and expansion of educational opportunity in the United States.

More students are going to high school, fewer are dropping out, and more high school graduates are aspiring toward college. This means that those taking the SAT last year were different from those taking the test fifteen years ago, insofar as they included a new group that was less prepared for college. Similar conclusions have been reached by others who have studied the problem, such as L. A. Munday in "Declining Admission Test Scores" in the March, 1976 American College Testing Program Research Report, and Annegret Harnischfeger and David E. Wiley, "Achievement Test Scores Drop. So What?" in the March, 1976 Educational Researcher. We have had many indications for a long time regarding the underlying reasons for the decline in college admission examination scores.

The Wirtz panel carefully avoided the temptation of pronouncing a single cause or a single solution. It warned against a "lowering of standards" to accommodate the "special needs of new and unfamiliar students." Members of the panel called for "cool-headedness" in assessing the entire subject.

The intensity with which the Wirtz study has been read and the respect it has evoked have not quite extended to heeding the admonition to "cool-headedness." Unfortunately, people are still demanding scapegoats for the low test scores. Some blame our teachers. George Gallup told us in September, 1976, that many parents are blaming themselves. The National Education Association, in turn, blames the tests.

In its recently issued "Alternatives to Standardized Testing," which repudiates standardized tests, the NEA advised colleges to rely "on the personal interview, on high-school class standing using the professional judgment of instructors, on grades, on student products, on artistic achievement, and on recommendations from teachers." How these criteria are more predictive and equitable than standardized tests and how they may be applied to the million-and-a-half college aspirants who presently take the SAT examination each year, NEA does not reveal. In short, its alternative to college admission tests is "to give no tests at all."

There is little danger that the NEA proposition will be taken seriously. Carried to its logical conclusion, this posture would also lead us to do away with tests used at all levels of education to determine the need for compensatory education and remediation. The idea of abolishing college admission tests, because they are imperfect or because they report unwelcome facts, is nothing short of professional nihilism.

The responsive teacher union has an obligation to interact with the larger society. It cannot avoid developing strategies to change misperceptions about our profession. The public is confused about testing.

Those who hear that "Johnny can't read" also conclude that he is completely illiterate. At best, a marginal recognition exists of the contemporary idea of functional illiteracy -- the concept that there are degrees of ability when it comes to reading, writing, and mathematical computation. Unfortunately, the average person barely comprehends this concept, which is central to the entire controversy over the meaning of standardized test results. Most people approach the subject of testing by creating false dichotomies: either a student can or cannot perform at acceptable levels. Instead of proposing, therefore, to abolish all standardized tests because they are misunderstood or their information may be misused, the teacher unions have an obligation to educate the public about what tests really mean. We should not run away from the issues that legitimately concern everyone only for want of perfection in the testing process or because testing creates difficulties for teachers. We should not reinforce the attitude that teachers have something to hide. They do not.

Furthermore, to deny categorically the validity of standardized tests goes against the grain of a seldom-mentioned feature of American civilization. Daniel J. Boorstin reminded us in his prize-winning volume, The Americans: The Democratic Experience (1973), that the United States is a "statistical" society, which since its beginning as a republic has always had an intense respect for numbers. Indeed, considering the fact that the Japanese actually celebrate a national holiday known as "Statistics Day," one might conclude that a dependence on numbers is a pre-condition of modernity itself. Tests are an inseparable component of the culture of numbers, just as the statistical culture is inseparable from the fabric of a modern society.

Boorstin is most convincing about the importance of numbers in his intriguing appraisal of the "statistical communities" that make up the democratic United States. While our national fascination with data may be excessive, we cannot dismiss all data as irrelevant. To recommend a position that standardized tests should be abandoned, that we must work to deprive our citizens of the statistics generated by virtually all forms of educational testing, is an emotional rather than an empirical stragegy. It can deprive us of a valuable educational tool, to which there is no practical alternative, and can only alienate the teacher unions from the overwhelming majority of people who care about the quality of life in this country. The NEA posture on testing will be viewed with hostility by the lay public and rightfully so. It cannot serve the interest of teachers at a time of unprecedented challenges in education.

Teacher unions have responsibilities. They must not detach themselves from reform of the testing system just because the tests have imperfections. A broad program of education can be more effective in protecting our members than a narrow platform of negativism. Ultimately, the teacher unions have to retain the respect of the public if they are to serve their members and their professions.

As teachers and unionists, we are very much concerned with the issues raised here with regard to colleges and universities. For this reason we are thoroughly committed to universal access to postsecondary education. This is a responsible position in the face of the problems raised by the general application of college admission test scores. Such access would still require some means of determining which college the student should attend and which program of study the individual should be enrolled in. Contrary to the illusions of the NEA, open admissions (as experienced at The City University of New York) does not do away with the need for tests. In fact, it intensifies the need for tests to identify the handicaps of entering students and to provide for their remediation. With universal access, however, the terrible burden now carried by admission tests would be eased and the penalty of error would be reduced for the student.

But universal access is not a reality now, and until it becomes a reality, we will continue to be confronted by the imperfection of admission tests and the indeterminate victimization of students by that imperfection. This presentation does not attempt to evaluate the evaluators. We have seen no responsible challenge to the general endorsement [illegible]en by the Wirtz panel to the validity of SAT scores and their predictive value for all segments of our society. One might note, in fact, that the universities of the People's Republic of China are returning to the college entrance examination requirement and that the University of California last month adopted a new admission policy that gives greater weight to standardized test scores.

The solution, then, must lie in the ultimate source of the problem, the uneven quality of educational opportunity available to youngsters for reasons that are beyond the control of educational institutions and teachers. The United States Supreme Court recognized this reality in its decision last June upholding a court-ordered desegregation plan for Detroit. The lower court held that ordering integration was insufficient for those students suffering from the negative effects of segregated schooling beyond the City's culpability and responsibility. It required the state of Michigan to pay part of the cost of the remedial programs called for in the plan, so that those entering integrated schools may be helped to overcome previous disabilities in the instruction and facilities they received. In what other way could the mandate for remediation have been possible except through standardized tests?

If we acknowledge the differences in the schooling made available to our elementary and high school pupils, we can then work toward improving the quality and equality of secondary school education so that all students receive the best possible preparation for postsecondary education. We must, at the same time, find and remove whatever cultural or sexual bias exists in the college admission test instruments. We should remove the barriers to access to postsecondary education, so that the standardized tests which indicate the quality of preparation for college may be relegated to the diagnostic role of placement tests rather than admission tests.

Teacher unions cannot appear to dismiss merit by proposing to eliminate all standardized tests and remain true to their profession. Just as merit should continue to be the criterion for certifying, evaluating, retaining, and promoting teachers, merit should also be the criterion for evaluating students. There is no other position for the responsible union. If tests have been the harbinger of bad news, we must confront that news and solve the problems it alerts us to rather than lynch the harbinger.

TESTING AND TEACHER UNIONS

Robert L. Ebel
Professor of Educational Psychology
Michigan State University

For purposes of this presentation the National Education Association, the various affiliated state education associations, the American Association of University Professors, and, of course, the American Federation of Teachers, will be regarded as teacher unions. All of these, save perhaps the last, were once and may still claim to be, professional associations. But clearly their goals and activities have shifted from that of exchanging ideas intended to improve the professional excellence of individual members to that of collective action to improve salaries and working conditions. No doubt the shift was inevitable in the evolution of our industrial society and democratic form of government. In some respects the shift has been beneficial. But in others it has obviously been detrimental.

Living teaches hard lessons that we are sometimes slow to learn. There is a price to be paid for every improvement we undertake. Every advance has its cost. That is as true of the unionization of teachers as is of the release of nuclear energy or the development of phosphate detergents. Success in one area brings problems in other areas. Unionization of teachers lessened their motivation and rewards for excellence in teaching and lessened the teacher's sense of individual responsibility for the achievements of pupils.

What has all this to do with testing? Testing is a means of assessing the competence of individuals; the competence of individual teachers to teach, the competence of the individual pupils they have taught. Unions, on the other hand, are not primarily concerned with competence, but with welfare, not with unique individuals, but with homogeneous groups. Unions are most directly concerned with salaries, job security, fringe benefits, conditions of work. Competence and welfare are both important goals, but they are not identical goals. In the days before teacher unions and uniform salary schedules, there was a substantial correlation between competence and welfare. Now it is much less. Excellence in teaching is of no great benefit to a member of a teacher union. It may be a handicap. Outstanding performance by one can lead to higher expectations from all, and the "all" do not much like that.

Thus it is not at all surprising that testing of the competence of teachers to teach or of the success of their pupils in learning is strongly opposed by some teacher unions. Testing is incompatible with the purposes of the union because it is more likely to threaten than to protect its welfare. Leadership of a union that advocated testing of the competence of its members to do the work they are employed to do or testing of the quality of the work they have done (i.e., the success of their pupils in learning) would have to be judged far less sane than most union leaders in fact are.

But union leaders who oppose testing seldom explain clearly why they are, and necessarily must be, opposed to testing. They seldom point to the inherent conflict between individual efforts to achieve excellence and group efforts to achieve power (and to protect every member almost regardless of competence). Instead, they criticize the adequacy of the tests and the influence of testing on pupils. They pose as competent professional judges of these matters, never acknowledging their conflict of interest and the bias it produces. The danger to education and hence to our society is that the public may take these criticisms at face value as the impartial judgments of well-qualified professionals. This they quite often are not. Their biases need to be recognized clearly. The validity of their criticisms need to be examined closely.

Tests:

1. are invalid, not really testing what they are intended to test.
2. are poor defined in terms of what they are intended to test.
3. only test the pupil's memory for trivial factual details.
4. are limited to testing knowledge, ignoring important affective outcomes of education.
5. measure mainly a pupil's skill in taking tests, not his possession of useful knowledge.
6. are biased against cultural minorities.

What are the major complaints concerning the influence of testing on education?

Testing:

1. induces teachers to teach to the tests.
2. imposes external controls on local school curricula.
3. fosters competition and cheating instead of cooperation.
4. upsets students, makes them anxious and teaches them to hate schooling.

Let us consider briefly each of these criticisms and complaints.

First, validity. What the test tests directly is the pupils success in completing the tasks set for him by the test questions; spelling a word, answering questions about a paragraph, solving a problem, etc. Usually the test tasks are a sample from a domain of similar tasks that the pupil might encounter later on in life and that it would be useful for him to be able to handle successfully. We make the reasonable inference that a pupil's degree of success on the test tasks provides a reasonable estimate of his probable success with other tasks from the domain. We also infer that what a pupil can or can not do on the test will give a reasonable indication of what the pupil can do on similar tasks outside the test. If these inferences are justifiable, as they almost always are, the test is indeed a valid measure of what it is intended to measure.

Next, defining what is to be measured. The domain from which the tests tasks are to be selected can be defined as clearly and in as much detail as may be desired. If it is poorly defined, the fault belongs not to testing in general but to some test developer in particular. In recent years a new type of test, referred to as an objectives-referenced test, has been proposed as a preferable alternative to the widely used domain sampling tests. It is true that the tasks included in an objectives-referenced test are defined more specifically than they are in a domain sampling test. But the area of learning preceding the test is correspondingly narrowed and sharply focused. In some areas of learning, particularly in the early grades, this narrowing and focus may be desirable.

For most other subjects of study, probably not. One does not need objectives-referencing to define clearly what a test is intended to measure.

Do tests require no more than memory for trivial factual details? Some do. None need to. If they do, it is the fault of the test constructor. Any test item can, and most do, require the examinee to recall the relevant knowledge he has gained and then to apply it to the novel question or problem in the test item. Good test items require the use of knowledge, not just the recall of information. Test specialists know this. Less experienced critics of testing do not or profess not to believe it.

Are paper and pencil tests limited to testing for knowledge? In general, they are, but in view of the great importance of knowledge in human affairs, in view of its great contribution to wisdom and virtue, in view of the amount of time good teachers spend in cultivating it, that limitation should trouble no one. Knowledge is important, but so are attitudes, interests, and values. If a teacher sets out deliberately to influence the affective characteristics of his pupils, he will almost certainly use knowledge as the principal means to the end he is seeking. He may also use rules of behavior enforced by rewards or punishments. Quite undeliberately and unintentionally he will influence the pupils affectively by the model of wisdom, decency, and humanity he provides for them. But where purposeful teaching is concerned, knowledge is the main thing. Tests which assess knowledge will do most of the job that needs to be done.

What of test taking skills? The claim that objective tests measure little more than test taking skills, that one who possesses such skills can do well on any objective test, regardless of the subject matter content of the test, is an absurd exaggeration. What is in fact true is that an examinee who is seriously deficient in elementary test taking skills, does not pay attention to the test directions, does not know how to mark his choices on the answer sheet, does not venture an answer unless certain that it is correct, wastes time on puzzling questions, such an examinee can impair his test score, perhaps seriously in some cases. But test taking skills are easy to learn and are quickly learned by most pupils. They are not, as some test critics have suggested, the special gifts of a favored few.

Test taking skills never enable an examinee to inflate his score on a good test. They are no substitute for competence. It is only the lack of such skills that can be harmful. Good teachers will help any pupils who lack such skills to acquire them as soon as possible.

Are tests biased against cultural minorities? They are not biased if they are achievement tests, the tests are administered fairly, and the scores are understood to measure only how well the examinee can handle tasks similar to those included in the test. When members of one group score lower than members of another group on a test of achievement, it is because the test is biased against them or is it becuase they have actually achieved less of the ability being tested, possibly through no fault of their own? On the other hand, if a minority child, who does not try as hard as non-minority classmates to do well on the test, scores lower than they do, is it the test that is biased or is it the test taker?

Aptitude or intelligence tests are another story. They involve predictions of future developments from measures of current abilities. If past opportunities to learn have been seriously limited, it is a grave error to assume that a person's current status is a good indication of that person's potential for learning. But note well that the bias here occurs in the prediction, not in the measurement of current status. The charge that tests of educational achievement are biased against cultural minorities, like most other charges against such tests, rests on no firm foundation.

Let us turn now to consider complaints concerning possible harmful effects of testing on education. Does testing induce teachers to "teach to the test"? It may do so in ways that are educationally beneficial. It may also do so in ways that are educationally harmful. What is the difference?

The beneficial way of teaching to a test is to help pupils to learn how to handle tasks similar to those included in the test. If it is a good test those tasks are a sample of a much larger domain of tasks which pupils need to be able to handle, in school or out. The harmful way is to give pupils advance information on specific questions that will appear on the test and to drill pupils on answers to those specific questions. Such teaching results in only superficial and ephemeral learning. Any teacher who claims that testing forces him to teach in this unproductive manner may be very foolish, very flabby ethically, or both.

Does testing impose external controls on local school curricula? To some degree it surely does. Is this bad? In most cases surely not. Ideal local curricula must necessarily include many elements that are common across all localities. Regardless of where we live our needs, aspirations, problems and prospects are much the same. And these days we do move about from one locality to another. Considerable uniformity in the curricula of schools in different localities is not only inevitable but also advantageous.

The ideal local curricula should also include some elements unique to a particular locality or even to a particular teacher, for some learning opportunities and some teaching interests and talents are indeed unique. The value of that uniqueness ought not to be allowed to dominate the curriculum. If doing well on external tests is regarded as one very important goal of a local school, the external control is likely to be beneficial rather than harmful.

Does testing harm children by fostering competition among them? Probably very rarely and certainly not necessarily. Testing permits comparisons between pupils in achievement, but comparisons may or may not involve competition. Further, competition in the classroom is quite different from that in the prize ring, where competitors aim to hurt each other or in business, where competitors try to take business away from each other. Those who compete in the classroom are all likely to gain from it, learning more than they would in the absence of competition. They may even cooperate in learning for their mutual benefit. If a student seeks to excel in learning by making it difficult for others to learn, shall we blame the goal -- excellence in learning -- or the student for using unfair means? Further, to blame cheating on competition to achieve excellence is like blaming banks for bank robberies.

Finally, consider the influence of testing on a child's anxieties and self-concept. The incidence of this alleged harm is less frequent and less serious than some critics would have us believe. Most teachers of young children are more concerned with pupils who do not care enough about how well they do on the test than with the rare pupil who seems to care too much. The child who, as is sometimes reported, breaks down in tears at the prospect of a test, has problems of security and adjustment that testing did not create and that abolition of tests can not cure. Being occasionally anxious is part of the price we pay for being human. Learning to cope with anxieties is part of the process of growing up.

Testing can augment or it can erode one aspect of the self-concept of a pupil. For most of us there are many such aspects. Deficiency in one area can be offset by excellence in another. A realistic self-concept is more to be desired than one of uncritical self-esteem or a hypercritical self-abasement. Justifiable self-esteem must be demonstrated, and tests provide one way of demonstrating it. If tests occasionally threaten one aspect of a pupil's self-concept, is this not likely to be therapeutic? It can lead the pupil to mend the shortcoming or it can lead to emphasis on some other avenue to excellence.

These have been brief answers to some of the most frequent criticisms of testing in the public schools. Such answers are unlikely to satisfy the critics, but they may suggest that there is much more to be said in favor of testing than the critics will allow. Standardized tests are by no means perfect, but the best of them today are very good indeed. They reflect the concensus of the best teachers on what is important for pupils to learn. They reflect the skill of competent test specialists and the experience they have gained in half a century of effort to improve the measurement of educational achievement. They provide meaningful, dependable measures of achievement in learning. None of the alternatives, NEA and other critics have to offer to standardized tests, can claim nearly as much perfection, precision, or dependability.

One of these alternatives deserves special comment. It is the objectives-referenced test which NEA has suggested tentatively and hesitantly as a possible replacement for the "worthless" norm-referenced standardized tests. An objectives-referenced test is intended to identify the different things a pupil has or has not learned in a particular subject of study. A norm-referenced test is intended to measure the amount a pupil has learned by comparing it with what other pupils have learned.

Actually there are more points of similarity than of difference between the two kinds of tests. The same question can be used in each, though those chosen for an objectives-referenced test are likely to be somewhat easier (every pupil is expected to be able to answer every question correctly) and to be concentrated on fewer objectives for learning.

Rules and suggestions for preparing good objectives-referenced tests are virtually identical with those for preparing good norm-referenced or standardized tests. Norms can be provided for an objectives-referenced test (though NEA regards this as a perversion to be carefully avoided). Objectives-referenced inferences can be made from pupil responses to the questions on a standardized test.

Standardized tests provide common yardsticks for measuring the amounts of learning pupils have achieved in different classes and in different schools. It is this potential for providing comparative data that seems to cause most of the opposition to such tests. Yet if we are interested in excellence in education, in stimulating and rewarding efforts to achieve excellence, comparisons can not be avoided. For excellence, like deficiency, is essentially a relative term.

It is difficult to accept the NEA contention that measures of student learning should not be used in the evaluation of teaching. Schools are for learning, and the job of the teacher is to facilitate that learning. How much is learned in any school setting depends, in part, on the skill and dedication of the teacher. It depends also, in part, on the abilities and efforts of the student. Of course, scores on standardized tests can not provide all of the evidence needed to judge a teacher's effectiveness. But to claim that test scores have nothing to contribute to such judgments is simply absurd.

From scores on college admissions tests, National Assessment testing, and other standardized tests, we have recently been getting some bad news about the educational achievements of our young people. The reaction of NEA is similar to that of ancient despots. If a messenger brings bad news, get rid of the messenger.

The antipathy of NEA for standardized tests is not shared by the public at large, nor by their elected representatives in government. In state after state legislatures have mandated programs of standardized testing to assess the effectiveness of instruction. There is currently less public confidence in the quality of education our young people are receiving than at any time in this century. For NEA to continue to oppose assessments of quality in education, identifies that organization as part of the problem, not of its solution.

Teacher unions have the important functions of promoting and defending the welfare of teachers. They were not designed and are not well equipped to promote excellence in teaching and learning. That responsibility belongs to the school administration, acting under the authority of the school board, the government, and ultimately the public at large. The public should not delegate to teacher unions the authority and responsibility for managing the enterprise of education, for setting goals, for assessing accomplishments. The public should not allow teacher unions to suppress, or even to suspend temporarily, the testing of teacher and pupil achievements in learning.

PART IV - TESTING USES AND MISUSES IN EDUCATION

TESTING USES AND MISUSES IN EDUCATION AS RELATED TO MINIMAL COMPETENCY TESTING

Thomas W. Mann
Director, APL Program
American College Testing

Of all the current issues on testing, the uses and misuses of testing, particularly as related to minimal competency assessment, is a key concern to educators. Teachers, high school administrators, members of local boards of education, and state legislators are relating competency assessment to accountability of the educational program. Parents and employers are also concerned. They want some assurance that a high school diploma, rather than merely being a certificate of accumulated credits and attendance, signifies that the holder possesses the basic skills necessary to cope in today's world. "Proficiency in the basic skills, statewide assessment of minimal competency, minimal achievement level testing, or tests of survival skills -- no matter how it is described, legislative and public interest in this issue will continue to grow." (Pipho, 1977).

No movement has affected education so rapidly and so forcefully as minimal competency testing or assessment. Twenty-six states have either legislation or state board of education rulings relating to minimal competency testing at the secondary level. The only commonality of legislative and state board of education efforts is the relationship to minimal competency assessment. No two mandates are the same; the efforts are as varied as states participating in this effort to provide minimal competency assessment. Some states, such as New Mexico, North Carolina, and Florida, have taken a strong, centralized approach. Other states, such as Oregon, Nebraska, and Colorado, give more freedom to local boards of education to make decisions relating to the identification of competencies and the establishment of standards.

The educational practitioner and lay leader alike have been inundated with proposals and counterproposals relating to the use and meaning of competency-based/performance-based instruments as a means of determining the functional ability of secondary and adult students. Thus the propriety of using criterion-referenced test instruments to determine levels of functional competency brings a new approach and a new set of problems related to assessment.

Which term -- competency/performance/proficiency testing -- best describes the minimal competency movement? The differences are in the criteria which are acceptable for minimal competency. What is more than adequate in one instance is unacceptable in another. The implication is that a series of real life competencies are identified which increase in complexity and scope as the student gains experience. Thus, what would be accepted as minimal would not be interpreted as ever becoming maximal competence or construed as being proficiency. Competency-based testing places the emphasis on a minimum standard.

"Performance-based," on the other hand, refers to the way in which individuals demonstrate knowledge and skill. A fundamental aspect involves the definition of competencies as indicators of successful performance in everyday, real-life activities.

The procedure of identifying and determining the number of competencies varies from local district identification in Oregon to identification of the goals and objectives by Adult Performance Level (APL) concept.

Development of the APL Concept

The APL materials are based upon research conducted at the University of Texas at Austin and supported by the United States Office of Education. The primary aims of that research were to specify the competencies needed for success in today's society and to develop devices for measuring such competencies in the adult population.

The Texas research team accomplished the first objective through the development of a general model of functional competency and the creation of a taxonomy of associated adult performances. This model may be described by the following four statements: 1) the term functional competence is meaningful only in a specific societal context; 2) functional competency is best described as the application of a set of skills to a set of general knowledge areas; 3) functional competency results from a combination of individual capabilities and societal requirements, and 4) functional competency is directly related to coping in adult life.

The Texas research team then defined objectives and performance indicators that were field-tested and revised until consensus about each objective was reached. In addition, each objective was validated against non-test measures of success (i.e., income, education, and occupational status).

National levels of functional competency were then determined. This involved the transformation of the performance indicators, each of which is related to a specific objective within a single content area, into a series of interview schedules. Interviews were then conducted with representative samples of the adult population. The resulting data were used to define levels of competency based on success as measured by income, job status, and education. Thus, an operational definition of functional competency and a conceptual framework for assessing competency were made available to the educational community.

The operational matrix with which the APL program functions is related to five life skills and five general content areas. A goal statement and a set of objectives have been developed for each of the five general content areas. Thus, the objectives become the criteria by which each assessment item measures performance on one of the five skills.

Figure 1

The Content-by-Skills Matrix. Examples of Tasks.

CONTENT AREAS → ↓ SKILLS	COMMUNITY RESOURSES	OCCUPATIONAL KNOWLEDGE	CONSUMER ECONOMICS	HEALTH	GOVERNMENT AND LAW
IDENTIFICATION OF FACTS AND TERMS	Knowing what a time zone is	Knowing what skills are needed for clerical jobs	Knowing what "bait and switch" is	Knowing what the normal human temperature is	Knowing what the Bill of Rights says
READING	Reading a bus schedule	Reading a want ad	Reading a contract	Reading a prescription label	Reading a ballot
WRITING	Writing a letter to make hotel reservations	Filling out a W-4 Form	Filing a consumer complaint	Answering a medical questionnaire	Writing a letter to a legislator
COMPUTATION	Computing a plane fare	Computing overtime earnings	Finding the best buy	Computing a daily dosage	Computing a statute of limitations
PROBLEM SOLVING	Determining where to go for help with a problem	Deciding what to say to a bother some co-worker	Deciding which of two decisions is better in economic terms	Deciding which meal is best, given a set of preconditions	Determining whether a given situation or action is legal

In order to understand APL Content Area Measure Level scores, one must understand the context within which these scores are interpreted. The initial APL research, conducted at the University of Texas, established the context. This context is a construct known as functional competency and a matrix of objectives and skills which identify sets of critical tasks. These tasks represent the consensus of adult educators, government officials, and others concerned with the functional competency of adults.

The APL Content Area Measures were developed by American College Testing (ACT) within that context to reflect mastery of those critical tasks. Furthermore, these measures were analyzed by independent content specialists who judged them to be content valid. Finally, a set of statistical procedures was applied which indicated that the items in the measures represent an adequate sample from the content domains.

Cutoffs for each of the Content Area Measures were set by consensus of adult education and measurement specialists to reflect three APL Levels:

0 - 50% correct	Level 1
51 - 75% correct	Level 2
76 - 100% correct	Level 3

The definitions of the three Levels are couched in instructional terms, as follows:

Level 1 - Has an inadequate degree of competency; a definite need for study and remediation to meet the APL goals and objectives through the application of basic skills.

Level 2 - Has a marginal degree of competency; a need for study and review to meet the APL goals and objectives through the application of basic skills.

Level 3 - Has an adequate degree of competency; may need some review to continue to meet the APL goals and objectives through the application of basic skills.

These definitions differ somewhat from the original APL Level definitions in that they are purely instructionally oriented. No attempt is made to predict success in life directly from Level scores on the APL Content Area Measures. By consensus of definition, Level scores are directly interpreted in terms of what individual learners have and have not mastered through the application of basic skills to the APL goals and objectives. To the educator, this means being able to determine more accurately those areas in which learners require remediation and/or instruction. This application is not unlike the typical application in the classroom where the teachers make judgments.

Minimal competency testing or assessment represents a significant trend which will continue to have a major impact upon our education system. Valid use of competency testing can provide state and local school systems with needed data which, when properly interpreted, can establish a means of accountability, meaningful to parents, teachers, administrators, and the American citizenry.

Let me indicate four specific and valid uses of competency testing:

1. Provide a general assessment of the educational program meeting the needs of students to function as productive citizens.

Minimal competency testing has the potential to make our school systems more responsive to those most vitally concerned -- parents and members of the community. Properly used minimal competency testing can provide a new perspective for determining how well the educational system is preparing the student to function competently in today's society.

2. Provide concrete and specific information to students and parents relating to the ability of students to function as productive citizens.

If minimal competency is defined as a level of educational performance which predicts a reasonably good probability of a satisfactory, productive life, it seems reasonable that the following criteria should be used in establishing a minimal competency level:

a) The performance level should be above that which is achieved by "dependent citizens".

b) There should be evidence that most citizens scoring at this level can live productive lives.

3) Establish a pragmatic and realistic expectation for parents and teachers in terms of what students should be able to do once having mastered the basic skills.

Individuals need to be competent in the skills necessary to cope at a level sufficient to learn job specifications or to pursue the requirements to enter postsecondary education. A performance below the minimum level (and denial of high school diploma) indicates:

a) That the probability for a student living a satisfying and productive life is reduced.

b) That the student and his family should realistically reassess his aptitudes and attitudes in order to help improve his performance level and his future prospects.

c) That in some situations, the school will not award diplomas, because it would mislead the students regarding their readiness for adult responsibilities.

4) Enhance the credibility for the awarding of a high school diploma.

Credibility can best be achieved through the participation of citizens in public hearings and elective processes in the adoption of minimal competency objectives. These would be ideally set at the local level but could also be adopted at the state level with provision for addition of local objectives (revisions and deletions would lead to chaos). Once these objectives had been adopted - and they could indeed reflect esoteric, aesthetic, moral, affective, and other kinds of objectives -- a credential reflecting mastery of these minimum competencies would carry weight.

Tests are tools, not magic wands. Even with future development of minimal competency tests, there can be misuses.

Misuse of a minimal competency instrument can:

1. Establish a criterion wrongly based on student performance.

Implementation and use of tests and setting of standards must be carefully documented, especially if those standards are applied without opportunity for individuals to acquire skills. It is a misuse to determine standards and not do something about it. Minimal competency is as much a social issue (and possibly more so) as an education issue.

2. Create false expectations of students on cutoff criteria if established too low.

Focusing on test and person, rather than instruction and system, i.e., using test to punish students rather than to upgrade education, is not a valid use.

3. Used to assess the total scope of an individual's education program and used to indicate achievement in the humanities or esoteric content areas.

Adoption of test and cutoffs without thorough examination of constructs and objectives by professional and lay persons concerned with education in a given community is not a valid use of tests.

Conclusion:

A major new thrust which promises vigor and innovativeness is being applied to testing and assessment. This involves dealing with all the elements which comprise total educational systems. Minimal competency assessment offers a process for directly focusing on functional literacy and the coping skills necessary to meet minimum societal demands, the results of which can provide new meaning and stability to secondary education and credentialing.

ON THE USES AND MISUSES OF STANDARDIZED TESTS IN EDUCATION

E. Belvin Williams
Senior Vice President
Educational Testing Service

If one were only to view the major legal arena and the decisions reached therein over the last decade, the concern about proper test use would most likely appear to be of recent origin. Litigious encounters have been numerous, resulting in legal landmarks in _Greggs v. Duke Power_, _Albermarle Paper Co. v. Moody_, _Tyler v. Vickery_, and _Washington v. Davis_. But as Thorndike and Hagen (1961) point out, both psychological and educational measurement have been subjected to marked criticism since approximately 1930. They remark that the criticisms have been either"...of (a) the incompleteness and imperfection of the information yielded by our measurement procedures or (b) the unwise things that we do with that information (p. 13)."

To be sure, professional interest in test construction, administration, and interpretation spans almost a half century. Recognized and widely accepted guidelines regarding the construction and publication of tests were developed almost a quarter of a century ago (American Psychological Association, 1954) and have received periodic review and revision (American Psychological Association, 1966, 1974). Despite the early effort, professional sensitivity to issues of test misuse was not as sharp 25 years ago as it is today. Witness the absence of indexed reference to "test use" in _Educational Measurement_ (1951) when the '54 guidelines were already under discussion, relative to that of _Educational_

Measurement (1971). Thus, for both the professional and the layman, there has been an increasing awareness of problems generated by casual or uncritical test use. Fundamental to the rise in popularity of both psychological and educational tests has been the heightened "decision demand." As expressed by Thorndike and Hagen (1961)

> the educator or the practical psychologist is continually faced with the necessity of arriving at some decision as to a course of action. He must decide what to do about an individual or individuals, or he must help the person himself decide what to do.
>
> Our basic assumption is that sound decisions arise out of relevant knowledge of the individual or individuals.
>
> We assume basically that knowledge is good, that knowledge is power, that knowledge is the basis for effective control of the problems that confront us from day to day. This is a basic tenet of our faith (pp. 7-8).

If this tenet of faith is accepted, it can then be asked whether "test misuse" has only social and ethical meaning, or whether it has a technical definition as well, subordinated to the more encompassing universals of ethical behavior. In the first instance, one might ask whether in a world of perfect instrumentation, e.g., no error of measurement, there can be anything other than social or ethical "test misuse." In the second instance, one necessarily asks whether, in a world of imperfect instrumentation, e.g., acknowledged error of measurement, there are only the usual parameters defining errors of inference. Hence, there is not a technical meaning to the notion "test misuse," but only errors of inference which should be guarded against as effectively as possible. This query is not academic hair splitting. The answer or answers direct attention to the burdens of responsibility and to corrective strategies.

For purposes of argument, one can readily concede that, in a world of perfect instrumentation, uses of tests are conceivable which would be in conflict with some closely held set of social values. Only the limits of human imagination and benevolence would prevent such a conflict. In a world of imperfect instrumentation, whether there can be technical "test misuse" or only oversight of valid inference, is not so clear. A few instances bear further attention.

Recalling the basic tenet of faith set forth by Thorndike and Hagen (1961), one must ask how it is ascertained that relevant knowledge has been gained. With respect to tests, it is usually said that the confirmation of relevance is gained through the validation of the test. Commenting on this phrase Cronback (1971) argues that

> one validates, not a test, but an *interpretation of data arising from a specified procedure*.
>
> Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test is valid (p. 447).

Given Cronback's formulation, one can conceive of situations in which judgments based on interpretations of data may be misapplied, should two or more specific validation procedures be confused. Consider the different procedures that pertain to criterion-referenced relative to norm-referenced testing. In the latter case, the interest is to gain reliable discrimination between any two individuals at arbitrary points on the scale; in the former, the overriding interest is to gain reliable discrimination only at a single point on the scale. Given the different objectives, the test construction and psychometric estimates are likely to be based on quite different procedures. Hence, to judge two persons as different when both are above or below the criterion points, but do not have the same score, might be said to be a misuse of technical information.

But even if judgments on validated interpretations were appropriately related to specific procedures, one can ask whether use of the test at all could constitute test misuse. The situation identified by Meehl and Rosen (1955) is frequently encountered by the potential user of some psychometric instruments:

> The efficiency of the great majority of psychometric devices reported in the clinical psychology literature is difficult or impossible to evaluate for the following reason:
>
> a. Base rates are virtually never reported. It is, therefore, difficult to determine whether or not a given device results in a greater number of correct decisions than would be possible solely on the basis of the rates from previous experience. When, however, the base rates can be estimated, the reported claims of efficiency of psychometric instruments are often seen to be without foundation (p. 194).

Meehl and Rosen demonstrated the non-utility of test use where the occurrence of the phenomenon, e.g., poor adjustment, good adjustment or college success in the real world, differs markedly from the test's ability to detect or reflect the phenomenon faithfully. Indeed, fewer correct classification decisions could be made with the test than without it.

In more instances than one would care to list, test data are too heavily weighted. Thorndike (1963) calls attention to certain misuses frequently noted in educational settings. Clearly, the concept of overachievement may in fact be undergirded by the uncritical acceptance of errors in prediction. The same is true for the concept of underachievement. Psychometric limits and realities have in numerous instances become significant bases for social and educational decision making.

Emphasis of the limitation of tests such as those used in college admission decisions does not guarantee correct use. It is well known that test scores or the combination of test scores and grade point averages from high school are insufficient to account for all of the determinants of college success (Angoff, 1970). Yet, evidence is scarce to support the view that decision makers either collect or evaluate in a systematic way the contributions of other data sets in reaching admissions decisions.

Even if the tests were used with proper caution, the problem of fair test use as identified by Thorndike (1971) and Darlington (1971) would still remain. As pointed out by both of these authors and further articulated by Gross and Su (1975), and Petersen and Novick (1976), integral to the technical problem is a value decision that must be made by someone on someone's behalf and probably imperfectly, regardless of the amount of data or information available.

The problem of test misuse is inextricably intertwined with specific technical limitations of tests and the general limits of all empirical inquiries. For this reason, the corrective steps indicated for curtailing test misuse are varied. Clearly, test publishers have a major responsibility in such efforts. They must set forth lucidly and precisely the limitations of their instruments. At the very minimum, the burden of providing relevant information about testing objectives, validation strategies, and interpretive constraints lie with test publishers for their particular instruments. Conjointly, there must be almost unilateral effort on the part of test publishers to investigate possible sources of error that are both controllable and uncontrollable. Professionals -- psychologists, educators, lawyers or others -- certainly must critically question, in constructive and knowledgeable ways, the uses and possible misuses of test instruments.

The individual test taker, too, has a right and an obligation, as in all other things, to ask questions about his or her test experience and the intended use of test results. This is an important prerogative of citizenship in a democracy which ought not be carelessly regarded or infrequently used. Usually, the individual, whoever he or she is, and the institution, whatever it is, are served best when there is frank expression of interest and concern about the processes used to reach decisions, especially those having an impact on the pursuit of educational or professional goals.

The foregoing remarks are not to be construed as a devaluing of the systematic gathering, ordering, and interpreting of data on individuals. Efforts of a non-systematic sort usually are less defensible and potentially of greater harm than those exemplified in standardized testing procedures. But, the science of testing is not complete and our understanding of it is, at best, partial. We must, therefore, recognize our limitations and pursue the opportunities for increased knowledge.

In viewing testing as an inseparable interaction between instrument and user and in calling for increased alertness on the part of test maker and test taker, no suggestion is intended that the current problems will be automatically overcome. Quite the contrary. Increased vigilance may reveal moral dilemmas of an intractable sort.

Foremost among the conflicts that may be uncovered are those pitting different institutions against one another. Most likely test makers will have to increase the scope of their acknowledged responsibility to ascertain not only the authenticity of test results submitted by individuals but also the professional use of test scores by receiving institutions. Wherein lies institutional sovereignty, academic freedom, and institutional discretion? All of these concepts may require reexamination and redefinition.

With the rise of interest in competency testing, new and perplexing issues loom for test makers, test users, and the publics served. As McClung (1977) has pointed out, failure of educational institutions to make explicit their educational objectives, or failure to reflect educational objectives in curriculum construction ought not result in punishment of students by denying them graduation certificates. Educators must deliver what they promise. Tests must evaluate what has been delivered, not what has been promised. Students must be made aware of the offerings, as well as what they will be held responsible for knowing. The legal framework for close scrutiny of these interconnecting responsibilities is yet to be effected.

REFERENCES

American Psychological Association, American Educational Research Association, and National Co-ncil on Measurements Used in Education. Technical recommendations for psychological tests and diagnostics techniques. *Psychological Bulletin*, 1954, *51* (Suppl.)

American Psychological Association. *Standards for educational and psychological tests*. Washington: APA, 1974.

American Psychological Association. *Standards for educational and psychological tests and manuals*. Washington: APA, 1966.

Angoff, W. H. (Ed.) *The College Board technical manual: A description of research and development for the College Board Scholastic Aptitude Test*. Princeton, N.J.: Educational Testing Service, 1970.

Cronbach, R. J. Test validation. In R. L. Thorndike (Ed.) *Educational measurement*. Washington: American Council on Education, 1971, Pp. 443-507.

Darlington, R. B. Another look at "cultural fairness." *Journal of Educational Measurement*, 197 1, *8*, 71-82.

Gross, A. L. and Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology*, 1975, *60*, 345-351.

Lindquist, E. F. (Ed.) *Educational measurement*. Washington: American Council on Education, 1951.

McClung, M. S. Competency testing: potential for discrimination. *Clearinghouse Review*. September 1977, pp. 439-448.

Meehl, R. E. and Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 1955, *52*, 194-216.

Petersen, N. S. and Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, *13*, 3-29.

Thorndike, R. J., and Hagen, E. Measurement and evaluation in psychology and education. New York: Wiley, 1961.

Thorndike, R. L. (Ed.) Educational measurement. Washington: American Council on Education, 1971

Thorndike, R. L. The concepts of over-and underachievement. New York: Teachers College, Columbia University: 1963.

USE AND ABUSE OF TESTS IN EDUCATION

Elizabeth Hagen
Professor of Psychology and Education
Teachers College, Columbia University

We are living in a schizophrenic world as far as tests and testing are concerned. On the one hand, there are pressures from various groups to do away with all tests and testing in education, particularly standardized tests. These groups are extremely vocal in declaring that tests do not measure important educational outcomes, are culturally biased, deprive individuals of equal opportunity for advancement, and contribute nothing to the educational process. The public verbalization of these groups implies that education would be better and life would be fairer to all individuals if tests were banned.

On the other hand, 26 states have passed regulations requiring that all students pass a minimal competency test in order to obtain a high school diploma. Through the use of minimal competency tests, it is implied, and sometimes explicitly stated, the quality of education will be improved and schools will be successful in teaching all students to read, write and do mathematics at a level which will satisfy all critics of public schools.

Both points of view are extreme and the proponents of either are doomed to disappointment. The effects of inequalities of life in our society will not disappear by banning tests nor will significant increases in quality of education occur simply by giving tests. All groups involved in the present controversy over tests and quality of education need to listen and think more and talk or scream less if we are going to make any progress in solving the very real problems that face us.

Let me speculate for a while on what underlies the present controversy. First, there is, I think, genuine concern for certain segments of our population, particularly children, who are viewed as being short-changed in their life opportunities. The concern is shared by various groups: teachers, school administrators, college faculties, parents, the general lay public, and testers. No one of these groups has a "corner on the concern market," and each expresses its concerns in different ways.

Both educators and legislators have striven valiantly, in their different ways, to improve the life chances of certain segments of our population but with limited success. Although the limited success that we have had is disappointing, it should not be demeaned and belittled as much as it has been. It is frustrating to experiment, to work hard, to put your very soul on the line in your attempts to provide an optimum education for children, and then to evaluate for the effects of your efforts only to find little or no positive effect. It helps not at all to have to listen to critics, most of whom have never had to deal with the problems of educating all children not a select few, who not only are derisive of your attempts but also belittle educators using expressions such as lazy, stupid, self-centered, prejudiced and others equally pejorative. In such a climate of criticism and lack of understanding, it is no wonder that educators, particularly those in public schools, have attacked tests as being irrelevant, invalid and unreliable. After all, tests, particularly standardized tests, are placing their failures in public view. If we could get rid of the tests, we could get rid of the criticisms of schooling. Of course, we still would not be free from the basic problems of how to provide quality educational programs for all children. You can kill the messenger who brings bad news of the defeat in battle, but such an act does not turn the defeat into a victory.

Second, I think that much of the present controversy, both in education and in testing, arises from different value structures, different priorities in values, and different belief systems. It is a common human failing to think that one's own values, priorities and beliefs are not only the ones held by the majority of people but also are the "right" ones to hold. One only has to look at the arguments presented to the Supreme Court in the Bakke case to show how value systems differ. Those who believe in merit think that higher education opportunities should be open primarily to those that have earned them through their own efforts and achievements. People who value highly merit of this kind generally believe in tests, and tests have served well over the years to identify individuals with talent who come from social and economic backgrounds where such talent is likely to be overlooked.

On the other hand, a more egalitarian group of people exist who basically believe that anyone who wants the opportunity to go to college, medical school, law school, or any other type of school should be granted the opportunity. In some instances, for some types of postsecondary schools, one can provide for open admissions and permit anyone to enter. If one provides a place for everyone who wants to enter, the criticisms still continue, but usually they are concerned with standards of quality of education or failure of the educational institution to adapt teaching methods or procedures to maximize the probability of success of all entering students. Open-door admissions become revolving door admissions and the low-achieving entrant becomes the low-achieving "exiter."

Standardized tests do not usually enter into decision making in these institutions as far as admissions are concerned or in decisions about dropping a student. However, standardized tests are frequently used in open-admissions institutions for diagnostic and placement purposes. These tests frequently show that the reading and quantitative abilities of a significant portion of the entering students are at a much lower level than those needed to succeed in college-level courses in the way that such courses are usually taught. Placement, too frequently, becomes placement in remedial courses that are at the secondary level or lower, and the students in these courses frequently become discouraged because they rightly sense that they are not a part of the college. It should be obvious, however, that tests used for diagnosis did not make the decision about placement nor about the type of instruction that was provided. Tests never make decisions: they merely describe a performance on a certain set of tasks.

This point is important to keep in mind when we turn to a different situation in higher education where there are not enough places for everyone who wants to get in. This situation exists in almost all of our professional schools, such as medicine, law, and veterinary medicine. When the places are limited so that only one in ten or a hundred or a thousand can be admitted, how shall the choices be made? Everyone knows that to become a physician or lawyer or veterinarian one must first be able to complete successfully the educational requirements. If blacks, chicanos, and women do not gain entry to these types of professional schools, then their numbers in the professions will remain small. For some this would represent a social inequity that is intolerable. Some will argue, however, that since the number of places in professional schools is limited, the places should be awarded to those individuals who have the highest probability of completing the educational program.

The limited number of admissions brings our value systems to the forefront. Which set of values is going to be predominant -- those of merit or those of eqalitarianism? No matter which set of values is chosen, there are going to be perceived inequities in the distribution of these types of life opportunities. Tests play no role in these types of value decisions. Only people can make value decisions. However, when they do, they should have the courage to stand behind them and be willing to state the true bases on which decisions have been made. If one has made the decision that admission to an entering class should be done in such a way as to guarantee a certain composition according to race, ethnicity, or sex, tests can provide useful information, but not all of the information needed to select students. In this situation, though, one does not have a single applicant pool but instead has several applicant pools -- one of each subgroup that is to be represented in the entering class. If, on the other hand, people have decided that, when there are limited numbers of openings, they should go to individuals with the highest probability of completing the educational program, then there should be a single applicant pool. The choice then should be strictly on the basis of predictor scores.

Rarely, if ever, is there a single predictor used for admission to college or professional school; instead multiple predictors are used. The kinds of information used by admission committees come from records of past achievement in terms of grades, letters of recommendations or ratings by previous teachers or by other persons, interviews of candidates, tests, and frequently other sources too numerous and heterogeneous to mention. In these admission settings, it is relatively rare for tests to be given maximum weight in deciding who shall be admitted. Yet tests are frequently assigned the greatest share of the blame when a candidate fails to be admitted. To some extent, I think, this occurs because members of admission committees are somewhat cowardly in reporting the true reasons for refusing admission. To some extent, it occurs because admission decisions represent a complex mixture of subjective and objective opinion so that it becomes difficult to report specifically the reasons for rejection. If admission committees are pressed for reasons for failure to admit, a great tendency exists to make the test score the scapegoat. For someone in the measurement area, it is puzzling as to why test scores are the subject of attack more than are the interviews, which are much more subject to error, prejudice, and other idiosyncrasies of the interviewer than are tests. I guess most of us think we can impress and "con" another human being better than we can an inanimate test.

Let me turn, now, to the third factor that I think contributes to the controversy, namely our attitudes towards individual differences. On the one hand, we hear frequent verbalizations about the uniqueness of the individual, about treating each person as an individual, about individualizing instruction and about respecting and fostering individual differences. On the other hand, we hear furious attacks on any measurement procedure that reveals certain individual differences, such as those in reading ability, quantitative ability, or verbal ability. As one listens to or reads the attacks on tests for revealing these differences, one must conclude that some people think certain individual differences should be allowed to exist but others should not. For example, one hears no great outcry when individual differences in physical attributes, athletic prowess, or artistic and musical abilities are identified and rewarded. But, when one identifies individual differences in quantitative or verbal or general reasoning abilities, a surge of criticism follows. Why is this so?

First, I think it is because our society and our schools implicitly and explicitly assign a much higher prestige value to quantitative and general verbal abilities than they do to any other abilities. Our educational processes and methods at all levels are structured in a way that require all individuals to process information in an analytic, temporally-sequencing way using primarily verbal modalities of input and output. An eminent researcher on the brain has stated that we educate only one-half of the brain, the left hemisphere. We completely ignore the right hemisphere; therefore, we do not value the individual who processes information holistically. We do not know how to structure our educational activities to use other kinds of abilities. An individual who has not developed at least average verbal and quantitative abilities has a difficult time getting along in our educational settings from kindergarten on. As a result, these individuals find their life opportunities limited.

Second, many people believe that individual differences in verbal and quantitative abilities, as well as in spatio-visual and mechanical abilities, are artifacts of the tests and thus are not real. These are the people who claim the tests are culturally biased and reflect nothing except differences in life experiences. But these differences in life experiences are real; they affect the development of the individual. If a girl scores low on a mechanical ability test, one can claim that she scored low because our culture discourages girls from being interested in and engaging in mechanical kinds of activities. This might be true, but it does not negate the fact that, at this point in time, the girl performed poorly on the mechanical tasks presented on the test.

The notion that all individual differences revealed by tests are due to cultural biases in the tests is a persistent one. But are they? In a recent address in California, the Secretary of the Department of Health, Education and Welfare stated that a reading test designed for high school students contained a passage on the stock market. He indicated that this passage would be an example of cultural bias in tests for minority groups. But would it? Is the stock market familiar to most majority students? I do not think so. When we are trying to develop reading skills, what are we really trying to do? I think we are trying to develop skills that permit an individual to comprehend and extract information from any written material. We are not teaching the individual to read a particular passage that is familiar to him or her. It seems to me that the purpose of reading is to gain information from new material; therefore, any new material, as long as it is appropriate in terms of vocabulary, sentence structure, and complexity for a particular age or grade group, can be considered culturally fair.

There is the equally persistent notion that if we could just find the optimum learning environment and life environment for each individual, then individual differences would disappear. This again is false. If, by some magic, we could provide both of these, individuals would be physically and emotionally healthier than they are now; they would read better and have better verbal and quantitative abilities. However, under optimum conditions for each individual, individual differences would increase not decrease.

Third, I think there is objection to identifying individual differences because of what happens to individuals after the differences are identified. Too frequently, the actions that we take are not constructive. We place a child who is having difficulty learning in the regular classroom in a special class. After placing the child there, we treat the child in such a way that it is impossible for him or her to move back into the mainstream of education again. But again, this is not the fault of the test that identified the child's weaknesses. Even if one did away with the test, the same actions are likely to be taken. The basic problem is not really one of testing but of devising new and more effective interventions to remedy the deficits that we have discovered in children.

Fourth, I think there are objections to identifying individual differences in the verbal and quantitative areas because people confuse differences among groups and differences among individuals. As a group, low socioeconomic children and minority children tend to score lower on tests that measure individual differences than do higher socioeconomic children or majority children. These group differences are probably due to some extent to opportunities that are available to each group. Although many critics will claim that all of these differences, or most of them, are due to cultural biases on tests, I do not think that these are important contributors to the group differences that are being found. It is not that I do not think tests can be improved; they can. However, the tests do appraise abilities and achievements that are related to success in educational settings. Until the demands of the educational settings are changed, the tests will continue to provide useful information. We should, though, make every effort to decrease the differences in opportunities for each individual. I am not sure we know very much about how to do this, but we should continue to try. Meanwhile, we still need to use tests as one source of information to determine the effectiveness of our efforts.

Let me turn now to the last factor that I want to identify as contributing to the controversy in testing and education, that is the actions that are taken, presumably on the basis of test results. I have mentioned some of these previously. There is no doubt that some people have done stupid things and justified doing them by referring to test scores. Children have been placed in special tracks or special classes and, from that point on, have virtually been forgotten and neglected. But stupid things are done without tests. You may recall a study done by Rist and reported in the Harvard Education Review about ten years ago. In that study, the kindergarten teacher, by the end of the eighth day in kindergarten, had sorted the children and seated them at three different tables. She also treated them differently during the entire year and passed along her differential judgments about them to other teachers. These children, through the first three grades, were stigmatized by the judgments that the kindergarten teacher made on the eighth day of kindergarten. The kindergarten teacher's judgments appeared to have been influenced by the child's cleanliness, manners, and general behavior. Are these relevant to a child's ability to learn? I would say no. One would like to think that this is an exceptional and rare case, but it is not.

I would prefer to have both test data and previous teachers' assessments of a child if I were preparing to teach him or her. Doing away with tests will not do away with stupidity.

Some people say that test results not only lead to ineffective and sometimes destructive action but also that they feed the mills of prejudice. These people point to public statements that are made about the genetic inferiority of certain groups of people. Again, however, these kinds of statements have been made for thousands of years. Test results did not implant these ideas in people's minds, and banning tests will not reduce the number of such statements.

Although the present climate is hostile to both tests and to schools, I think both will survive. Those of us who have devoted our professional lives to testing and evaluation need to pay much more attention to translating test scores into constructive actions. Many of us have been too concerned with the predictive validity of tests and have been too little concerned with what test scores mean in terms of behavior and constructive actions to be taken to facilitate the development of the individual. We need to work more closely with teachers and educators to determine what information they need in order to make education more effective and help them to get this information in a form which is useful to them. I think that most of us in testing recognize the imperfections of our present instruments and are working toward improving them. Teachers, educators and the public also must understand that not only are tests not perfect but neither are other means of appraisal of individuals. In an imperfect world, one must use multiple sources of information in order to make good decisions. Good tests can contribute one source of information.

A SUMMARY STATEMENT

Ellis B. Page
Professor of Educational Psychology
University of Connecticut

When Dr. Bossone, long ago, offered me the opportunity to summarize the proceedings of this conference, it seemed a golden opportunity: Imagine having the last word on so many controversial topics! But a true summary, now that the time has arrived, is very difficult to make, with so many different perspectives presented, so many legitimate but opposite positions. On one point, I think we can all agree: The tone of the conference has been just right -- at once pointed, yet polite; at times political, yet rational.

So far as the substance is concerned, let us consider seven issues, which are very much with us as we go our separate ways:

1) Is testing here to stay? On this, virtually everyone who spoke is in agreement. The alternative to standardized testing is, of course, *un*standardized testing. How do the criticisms of standardized testing sound when applied to *un*standardized testing? Is such testing also to be considered invalid, used with poorly defined goals, emphasizing factual details, measuring knowledge only, measuring skills in test-taking, and biased against minorities? Is there one of these objections which does not seem all the stronger against *un*standardized testing?

2) Must we put test scores in perspective in order to make rational decisions? Again, we seem to be in agreement that single test scores are not a sufficient basis for important judgments concerning individuals,

groups, or programs. Ways to combine information (for example, using operations research techniques) need far more dissemination and application.

3) Who should govern school testing programs? Teacher unions? Testing corporations? Federal courts? If federal judges are ill-equipped to govern testing, then how can they best select and listen to expert testimony?

4) What are the sources of test scores? Today serious students are agreed that strong environmental and genetic influences exist in individual IQ or achievement test performance. But we seem to be still quite ambiguous on the question of innate group differences. This is an area which both sides should resolve as affirmative action programs are more critically scrutinized by the judiciary.

5) How should tests be designed when aimed only at certain specific cut-points? Here we have great problems, for example, with diplomas to be awarded only on "competency." Most such competency cut-offs are primarily aimed at extremely low achievers, perhaps those who are 2.5 standard deviations from the mean. To achieve accuracy here, the tests should be designed with a good proportion of items in this level. Yet teachers will not be instructing their twelfth-grade classes in seventh-grade work! Thus the schools are caught in still another dilemma, for courts have ruled that tests are fair only if they reflect the instruction!

6) How should tests be used for low-achieving ethnic groups? General consensus exists among scientists that tests are not "biased" against minorities if we are willing to accept any identifiable objective measures

as the criterion of success. Yet the available criteria is challenged and the nature-nurture question looms large in the background.

7) Should we take seriously the idea of open admissions? Because so much of testing controversy swirls around the question of selective admissions, will the problem go away if we simply open all doors? Although this idea has been seriously presented here, most appear to agree that it is not a general solution. In the first place, who will so dictate to colleges and professional schools that they must take all students? In the second place does not such an "open admissions" policy simply lead to failure for those students who are unable to do college work?

For me, each of these issues has been sharpened and modified during these spirited two days. As we go our separate ways, I feel assured that while we do not all agree, we are more understanding of different points of view and more capable of playing our own respective roles in the world of testing.

SELECTED BIBLIOGRAPHY

American Psychological Association, "Standards for Educational and Psychological Tests and Manuals," Washington, D.C. 19966.

Angoff, W.H., "Scales, Norms, and Equivalent Scores," Educational Measurement (2nd Ed.), Washington, D.C.: American Council on Education, (1971), 508-600.

Beezer, Bruce G., "Testing: Educational Goals and Change," Intellect, 102, 2352, (November,1973), 114-15.

Bennett, Virginia C. and Bardon, Jack I., "Where Do We Go From Here?" Journal of School Psychology, 13, 4, (1975), 349-358.

Bossone, Richard M., Ed. and Troyka, Lynn Quitman, Ed., Proceedings The National Conference on Minimum Competencies: Trends and Issues. New York: City University of New York, 1977.

Brady, Elizabeth, "To Test or Not to Test," American Educator (AFT Journal), 1, 1, (Winter,1977), 3-9.

Campbell, J.T., "Tests are Valid for Minority Groups Too," Public Personnel Management, 2, (1973), 70-73.

Cantrell, R.P., Stenner, A.J. & Katzenmeyer, W.G., "Teacher Knowledge, Attitudes, and Classroom Teaching Correlates of Student Achievement," Journal of Educational Psychology, 69, (1977), 172-179.

Cleary, T.A., "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges," Journal of Educational Measurement, 5, (1968), 115-24.

Common Sense and Testing in English, Urbana, Illinois: The National Council of Teachers of English, 1975.

Crano, W.D., Kenny, D.A. & Campbell, D.T., "Does Intelligence Cause Achievement? A Cross-legged Panel Analysis," Journal of Educational Psychology, 63, (1972), 258-275.

Cronbach, Lee J., Essentials of Psychological Testing, 3rd Ed. New York: Harper and Row, 1970.

Day, James F., "Supreme Court Decides Hiring and Placement Procedures Must Be Valid," Journal of Employment Counseling, 9, 2, (June,1972), 74-77.

Dubois, P.H., A History of Psychological Testing, Boston: Allyn & Bacon, 1970.

Ebel, Robert L., "The Paradox of Educational Testing," NCME Measurement in Education, 7, 4, (Fall, 1976).

Ebel, Robert L., "What Do Educational Tests Test?" Educational Psychologist, 10, 2, (1973), 76-79.

Fiedler, William E., "The Employer's Burden of Proof in Employment Testing Cases," Journal of the College and University Personnel Association, 25, 3, (July/August,1974), 10-20.

Fincher, Cameron, "Personnel Testing and Public Policy," American Psychologist, 28, 6, (June,1973), 489-497.

Flygare, Thomas J., "The Hazelwood Case: How to Use Statistics to Prove (and Disprove) Hiring Discrimination," Phi Delta Kappan, 59, 3, (1977), 210-211.

Francis, Anne M., "Schooling, Testing, and Jobs: A Re-examination of Grigg, et al. v. Duke Power Company," NOLPE School Law Journal, 5, 1, (1975), 56-72.

Ghiselli, E.E., "The Validity of Aptitude Tests in Personnel Selection," Personnel Psychology, 26, (1973), 461-477.

Gorham, William A., Some Psychometric and Legal Considerations in Cooperative Testing. Washington, D.C.: Civil Service Commission, 1974.

Guilford, J.P., The Nature of Human Intelligence, New York: McGraw-Hill, 1967.

Grommon, Alfred H. (ed.), Reviews of Selected Published Tests in English, Urbana, Illinois: National Council of Teachers of English, 1976.

Holman, M.G. and Doctor, R., Educational and Psychological Testing, New York: Russell Sage Foundation, 1972.

Humphreys, L.G., "Statistical Definitions of Test Validity for Minority Groups," Journal of Applied Psychology, 58, (1973), 1-4.

Kamin, Leon J., "Social and Legal Consequences of I.Q. Tests as Classification Instruments: Some Warnings from our Past," Journal of School Psychology, 13, 4, (1975), 317-322.

Kay, Patricia M., "Fairness in Assessment for Personal Selection," Education and Urban Society, 8, 1, (November,1975), 63-72.

Kirkland, M.C., "The Effects of Tests on Students and Schools," Review of Educational Research, 41, (1971), 303-350.

Lehrhaupt, Arthur., "For Millions of Children, the United Teaching Profession Says No to Statewide Testing," NJEA Review, 46, 4, (December,1972), 24-5, 27.

Marlowe, John, "Testing, Testing ... Can You Hear Me?" Phi Delta Kappan, 58, 3, (1976), 256-257.

McKinney, J.D., Mason, J., Perkerson, K., and Clifford, M., "Relationship Between Classroom Behavior and Academic Achievement," Journal of Educational Psychology, 67, (1975), 198-203.

Millman, J., "Criterion-referenced Measurement," in Evaluation in Education, Berkeley, Calif.: McCutcham, 1974.

Miner, J., "Psychological Testing and Fair Employment Practice: A Testing Program that Does Not Discriminate," Personnel Psychology, 27, (1974), 49-62.

Morris, L.W. and Liebert, R.M., "Effects of Anxiery of Timed and Untimed Intelligence Tests: Another Look," Journal of Consulting and Clinical Psychology, 33, (1969), 240-244.

Munday, L., "Predicting College Grades in Predominantly Negro Colleges," Journal of Educational Measurement, 2, (1965), 157-160.

NAACP Report on Minority Testing, New York: National Association for the Advancement of Colored People Special Contribution Fund, 1976.

Neville, D., "The Relationship Between Reading and Intelligence Test Scores," The Reading Teacher, 18, (1965), 257-262.

Nolte, M. Chester, "Use and Misuse of Tests in Education: Legal Implications," Educational Horizons, 54, 1, (Fall,1975), 10-6.

Oakland, Thomas, "Assessing Minority Group Children: Chalenges for School Psychologists," Journal of School Psychology, 11, 4, (1973), 294-303.

Purkey, W.W., Self-Concept and School Achievement, Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Ratteray, Joan D., *The Testing of Cultural Groups. A Paradigmatic Analysis of the Literature on Testing and a Proposition*, California: Rand Corporation, 1974.

Richardson, K. and Spears, D., *Race and Intelligence: The Fallacies Behind the Race - IQ Controversy*, Baltimore: Penguin Books, 1972.

Robertson, David E., "Employment Testing and Discrimination," *Personnel Journal*, 54, 1, (January, 1975), 18-21, 56.

Rudman, Herbert C., "The Standardized Test Flap," *Phi Delta Kappan*, 59, 3, (1977), 179-185.

Schmidt, F. L. and Hunter, J. E., "Racial and Ethnic Bias in Tests: Divergent Implications of Two Definitions of Test Bias," *American Psychologist*, 29, (1974), 1-8.

"Some Things Parents Should Know About Testing," *Test Service Notebook 34*, Harcourt Brace, Jovanovich Inc., (1977), 1-4.

Stevenson, J., Parker, T., Williamson, A., Hegion, A., & Fish, E., "Predictive Value of Teachers' Ratings on Young Children," *Journal of Educational Psychology*, 68, (1976), 507-517.

Thorndike, Robert L., "Concepts of Culture Fairness," *Journal of Educational Measurement*, 8, (1971), 63-70.

Thorndike, Robert L. and Hagen, Elizabeth P., *Measurement and Evaluation in Psychology and Education*, 4th Ed., New York: John Wiley and Sons, 1977.

Turnbull, William, "A Partial View of ETS," *Phi Delta Kappan*, 59, 2, (1977), 103-106.

Wahlstrom, M., and Boersma, F. J., "The Influence of Test-wiseness Upon Achievement," *Educational and Psychological Measurement*, 28, (1968), 413-420.

Walden, John C., "The Courts Look at Standardized Testing," *National Elementary Principal*, 54, 4, (March-April, 1975), 80-81.

West, C. K. and Anderson, T. H., "The Question of Preponderant Causation in Teacher Expectation," *Review of Educational Research*, Fall, 1976.